# Large Multimodal Models as General In-Context Classifiers

## Supplementary Material

This supplementary material provides additional technical and implementation details, alongside extended experimental results and further analyses that complement the main paper. Specifically, the material is organized as follows:

- Sec. A provides details regarding the datasets.
- Sec. B extends the *closed-world* analysis from Tab. 1 (*Main*) by reporting granular, per-category performance breakdowns.
- Sec. C presents the full set of results for the *open-world* setting (Tab. 2 in *Main*), including evaluations on 4-shot and 8-shot configurations.
- Sec. D provides the complete data for the *streaming* experiments visualized in Fig. 6 (*Main*).
- Sec. E outlines the implementation specifics.
- Sec. F showcases an extended gallery of qualitative examples, complementing the visualizations in Fig. 4 (*Main*).

## A. Datasets

We summarize the evaluation datasets in Tab. 3. The experiments utilized the same training and test splits as previous work [10, 12].

Table 3. **Dataset details.** Summary details of the datasets used in our experiments.

| Abbr. | Dataset | Images | Classes |
|---|---|---|---|
| C101 | Caltech101 [14] | 2,465 | 100 |
| DTD | DTD [9] | 1,692 | 47 |
| ESAT | Eurosat [17] | 8,100 | 10 |
| FGVC | FGVC Aircraft [24] | 3,333 | 100 |
| FLWR | Flowers102 [27] | 2,463 | 102 |
| FOOD | Food101 [5] | 30,300 | 101 |
| PETS | Oxford Pets [28] | 3,669 | 37 |
| CARS | Stanford Cars [19] | 8,041 | 196 |
| S397 | SUN397 [38] | 19,850 | 397 |
| U101 | UCF101 [33] | 3,783 | 101 |

## B. Closed-world results

This section presents the detailed, per-dataset results of the experiments described in Sec. 3.3 and Tab. 1 (*Main*).

As they did not fit on a single page, we split the table into two parts, showing contrastive VLMs in Tab. 4a and generative LMMs in Tab. 4b.

For contrastive VLMs (Tab. 4a), we report the performance of the *Zero-Shot* model, Tip-Adapter [43] using the same backbone, and k-NN (as described in Sec. 3.3 in *Main*). For both Tip-Adapter and k-NN, we report results with 4, 8, and 16 shots. For each experiment, we report the average (*Avg.*) on the ten datasets, and highlight the delta $\Delta$ w.r.t. the average of the corresponding *Zero-Shot* model.

For LMMs (Tab. 4b), we report the performance of the *Vanilla* model (*i.e.*, multi-choice) and when using both a *Random* context and a *Similarity*-based context that retrieves images. *Similarity* ICL uses CLIP ViT-B/32 [30] to retrieve the most similar images from the few-shot pool. We report results in the 4-, 8-, and 16-shot settings.

The results for contrastive VLMs in Tab. 4a highlight the efficacy of adapter-based methods over simple nearest-neighbor retrieval in the few-shot regime. Tip-Adapter consistently outperforms the *Zero-Shot* baseline across all backbones and shot settings, achieving a peak improvement of +8.8% with ViT-B/16 at 16 shots. Conversely, the k-NN approach often struggles in low-shot scenarios (4 shots), frequently yielding negative deltas compared to the zero-shot baseline (*e.g.*, -8.4% for ViT-B/32). However, k-NN performance recovers as the number of support examples increases to 16. Overall, while scaling the backbone from ViT-B/32 to ViT-L/14 improves absolute accuracy metrics, the relative trends between Tip-Adapter and k-NN remain consistent.

In the case of generative LMMs (Tab. 4b), the data reveals a critical sensitivity to the quality of the in-context examples. The *Random* setting proves catastrophic across the board, causing massive performance degradation (*e.g.*, up to -48.7% for Qwen-2-VL), suggesting that irrelevant context acts as noise that confuses the model rather than aiding it. In contrast, *Similarity*-based retrieval unlocks significant performance gains. This is most notable in Phi-3.5-Vision, which scores +29.2% at 16 shots, and the Qwen family, where Qwen-2-VL achieves a +17.7% boost.

## C. Open-world results

This section details the results of our Open-World (OW) experiments. We conduct a comprehensive evaluation across all five models and ten datasets, testing three In-Context Learning (ICL) variants: *Random Context*, *Pseudo ICL*, and CIRCLE. For each method, we report performance at 4, 8, and 16 shots.

The detailed results are organized by model family and metric as follows:

- **Main results.** Tabs. 5a to 5c present the aggregated performance for the Qwen series, LLaVa OneVision, and the Phi series, respectively. These tables group results by dataset type: *Prototypical*, *Non-prototypical*, *Fine-grained*, and *Very fine-grained*.

Table 4. **Closed-world results.** Accuracy on the ten datasets. **Bold** indicates the best result for each VLM. Δ computed w.r.t. the Avg. of *Zero-Shot*.

(a) **Vision-Language Models (VLMs).**

| Model | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *CLIP ViT-B/32* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 92.4 | 44.6 | 45.1 | 19.2 | 69.3 | 80.5 | 87.3 | 60.1 | 62.5 | 64.5 | 62.6 | |
| TIP-Adapter | 4 | 93.3 | 49.3 | 57.5 | 21.8 | 74.8 | **80.7** | 87.9 | 63.0 | 64.9 | 67.1 | 66.0 | +3.4 |
| k-NN | 4 | 87.6 | 43.5 | 66.7 | 20.5 | 80.0 | 52.8 | 41.9 | 37.1 | 50.9 | 60.8 | 54.2 | -8.4 |
| TIP-Adapter | 8 | 93.4 | 52.4 | 63.4 | 23.5 | 79.0 | **80.7** | 88.3 | 65.4 | 66.4 | 69.2 | 68.2 | +5.6 |
| k-NN | 8 | 87.2 | 53.0 | 72.5 | 25.2 | 83.6 | 60.9 | 52.4 | 45.7 | 58.6 | 64.3 | 60.3 | -2.3 |
| TIP-Adapter | 16 | **93.9** | 56.1 | 65.2 | 25.6 | 82.8 | **80.7** | **88.5** | **67.8** | **68.8** | **71.2** | **70.1** | +7.5 |
| k-NN | 16 | 90.8 | **56.3** | **73.4** | **29.0** | **85.8** | 67.0 | 61.8 | 54.5 | 62.2 | 67.7 | 64.8 | +2.2 |
| *CLIP ViT-B/16* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 94.2 | 45.7 | 48.2 | 24.8 | 71.3 | 85.8 | 89.1 | 65.6 | 63.0 | 68.5 | 65.6 | |
| TIP-Adapter | 4 | 94.8 | 49.1 | 66.6 | 27.9 | 76.8 | 86.2 | 90.1 | 68.3 | 65.8 | 71.5 | 69.7 | +4.2 |
| k-NN | 4 | 90.2 | 46.6 | 72.7 | 29.5 | 86.9 | 65.6 | 58.0 | 50.2 | 53.6 | 62.6 | 61.6 | -4.0 |
| TIP-Adapter | 8 | 94.9 | 52.2 | 69.6 | 30.5 | 79.9 | 86.2 | 90.2 | 70.0 | 67.6 | 73.2 | 71.4 | +5.8 |
| k-NN | 8 | 90.6 | 56.3 | 73.7 | 32.0 | 90.3 | 71.8 | 65.2 | 59.3 | 60.6 | 68.3 | 66.8 | +1.2 |
| TIP-Adapter | 16 | **95.6** | 57.0 | **76.4** | 33.3 | 84.8 | **86.5** | **91.9** | **73.1** | **70.0** | **75.5** | **74.4** | +8.8 |
| k-NN | 16 | 93.1 | **58.2** | 76.2 | **36.3** | **92.0** | 77.0 | 75.3 | 65.7 | 64.4 | 70.6 | 70.9 | +5.3 |
| *CLIP ViT-L/14* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 96.8 | 53.7 | 60.2 | 32.7 | 80.8 | 91.0 | 93.5 | 76.8 | 68.0 | 75.8 | 72.9 | |
| TIP-Adapter | 4 | 97.1 | 57.3 | 68.8 | 37.1 | 84.9 | 91.2 | 93.8 | 79.0 | 70.1 | 77.0 | 75.6 | +2.7 |
| k-NN | 4 | 92.9 | 49.9 | 75.0 | 35.0 | 93.7 | 75.2 | 61.9 | 58.6 | 56.6 | 72.7 | 67.2 | -5.7 |
| TIP-Adapter | 8 | 97.3 | 59.9 | 74.5 | 40.4 | 87.9 | 91.3 | 94.0 | 80.2 | 71.4 | 78.1 | 77.5 | +4.6 |
| k-NN | 8 | 94.1 | 59.6 | 81.0 | 40.9 | 94.4 | 81.4 | 70.9 | 67.6 | 63.7 | 77.4 | 73.1 | +0.2 |
| TIP-Adapter | 16 | **97.5** | **64.4** | 77.7 | **44.3** | 92.2 | **91.4** | **94.2** | **82.8** | **73.6** | **80.1** | **79.8** | +6.9 |
| k-NN | 16 | 96.2 | 64.1 | **82.3** | 44.2 | **96.5** | 83.9 | 81.6 | 72.8 | 67.7 | 78.4 | 76.8 | +3.9 |

- **Metric-specific breakdowns.** We provide granular breakdowns for each metric across all ten datasets in the subsequent tables:
  - **Semantic Similarity** (SS) in Tabs. 6a to 6c.
  - **Concept Similarity** (bCS) in Tabs. 7a to 7c.
  - **Median Concept Similarity** (mCS) in Tabs. 8a to 8c.
  - **Llama Inclusion** (LI) in Tabs. 9a to 9c.

In all tables, we report the average score (*Avg.*) and the improvement (Δ) relative to the *Zero-Shot* baseline.

**Results discussion.** The open-world results (Tabs. 5a to 5c) demonstrate the consistent superiority of CIRCLE over both baselines and other ICL variants. Across all model families, constructing the context with our CIRCLE yields significant improvements in *Semantic Similarity* (SS), *Median Concept Similarity* (mCS), and *Llama Inclusion* (LI) compared to the *Zero-Shot* baseline. For instance, with Qwen2.5-VL, our 16-shot configuration improves SS on *Prototypical* datasets from 47.9 to 67.7, mCS from 31.1 to 67.2, and LI from 82.9 to 94.9. In contrast, *Random Context* acts as a distractor, causing performance degradation. This is particularly evident in LLaVa OneVision (Tab. 5b), where 16-shot *Random* context causes SS to lower from 56.2 (*Zero-Shot*) to 29.3, mCS from 53.4 to 29.6, and LI from 53.2 to 14.0, whereas CIRCLE recovers and boosts performance to 74.0, 74.0, and 72.2, respectively.

Furthermore, CIRCLE proves effective in handling fine-grained tasks, a setting where standard models typically struggle. In the *Very fine-grained* category, CIRCLE achieves consistent gains. Notably, Phi-3.5-Vision (Tab. 5c) sees its LI score nearly double, jumping from 54.2 in *Zero-Shot* to 99.6 with our CIRCLE in the default 16-shot setting. Similarly, Qwen2.5-VL improves from 69.0 to 93.6 in the same category. While *Pseudo ICL* generally offers some improvement over *Random*, it lacks the stability of our method, often underperforming the zero-shot baseline.

Finally, the results highlight the importance of scaling the number of shots when using relevant context. While 4-shot performance with CIRCLE already surpasses the *Zero-Shot* baseline in most metrics, such as the *Fine-grained* bCS score for Qwen2-VL rising from 62.9 to 66.4, the performance gap widens significantly at 16 shots (reaching 61.1 SS, 72.0 bCS, and 57.3 mCS). This behavior suggests that CIRCLE successfully leverages the additional diverse unlabeled images to construct a more informative context, a capability that is

Table 4. **Closed-world results.** Accuracy on the ten datasets. **Bold** indicates the best result for each LMM. Δ computed w.r.t. the average (Avg.) of *Vanilla*.

(b) **Large Multimodal Models (LMMs).**

| Model | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen-2-VL 7B* | | | | | | | | | | | | | |
| *Vanilla* | - | 91.6 | 61.9 | 33.9 | 50.3 | 69.8 | 82.3 | **86.8** | 16.4 | 48.9 | 70.9 | 61.3 | |
| *Random* | 4 | 71.5 | 41.7 | 21.0 | 11.1 | 20.0 | 64.1 | 55.3 | 7.0 | 10.8 | 37.4 | 34.0 | -27.3 |
| | 8 | 49.5 | 23.6 | 19.7 | 4.4 | 8.6 | 42.5 | 30.0 | 4.6 | 4.7 | 15.4 | 20.3 | -41.0 |
| | 16 | 20.6 | 16.9 | 17.5 | 2.3 | 3.9 | 39.0 | 15.5 | 2.7 | 1.2 | 5.9 | 12.6 | -48.7 |
| *Similarity* | 4 | 92.3 | 58.2 | 45.5 | 63.4 | 88.4 | 73.3 | 74.0 | 59.6 | 59.1 | 67.3 | 68.1 | +6.8 |
| | 8 | 93.8 | 66.9 | 46.8 | 67.8 | 92.7 | 79.2 | 75.5 | 74.2 | 67.2 | 73.5 | 73.8 | +12.5 |
| | 16 | **95.5** | **74.8** | **53.8** | **70.0** | **96.3** | **83.1** | 80.8 | **83.1** | **72.7** | **79.4** | **79.0** | +17.7 |
| *Qwen-2.5-VL 7B* | | | | | | | | | | | | | |
| *Vanilla* | - | 92.4 | 65.2 | 29.7 | 48.5 | 73.2 | **77.4** | **85.9** | 17.3 | 55.9 | 66.1 | 61.2 | |
| *Random* | 4 | 90.1 | 52.1 | 30.4 | 13.2 | 45.6 | 57.6 | 75.0 | 10.9 | 27.0 | 58.9 | 46.1 | -15.1 |
| | 8 | 88.6 | 51.9 | 30.8 | 12.1 | 48.3 | 53.8 | 72.2 | 11.9 | 28.2 | 58.3 | 45.6 | -15.6 |
| | 16 | 87.2 | 50.7 | 30.3 | 11.0 | 48.7 | 47.3 | 67.8 | 10.9 | 28.1 | 55.4 | 43.7 | -17.5 |
| *Similarity* | 4 | 92.2 | 50.2 | 38.6 | 50.4 | 90.4 | 63.4 | 72.5 | 34.5 | 59.0 | 61.9 | 61.3 | +0.1 |
| | 8 | 94.4 | 63.1 | 48.7 | 54.0 | 96.4 | 70.9 | 79.1 | 53.2 | 69.0 | 72.1 | 70.1 | +8.9 |
| | 16 | **95.6** | **72.5** | **55.6** | **60.8** | **98.3** | 76.9 | 85.2 | **69.0** | **74.0** | **76.8** | **76.5** | +15.3 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | | |
| *Vanilla* | - | 91.5 | **73.8** | 48.9 | **51.6** | 38.5 | **69.7** | **53.3** | 14.8 | 46.6 | **69.5** | 55.8 | |
| *Random* | 4 | 16.9 | 15.2 | 17.6 | 3.6 | 4.8 | 14.5 | 21.4 | 1.3 | 10.0 | 11.9 | 11.7 | -44.1 |
| | 8 | 20.3 | 18.0 | 22.5 | 4.7 | 5.8 | 14.7 | 19.3 | 1.3 | 10.8 | 15.7 | 13.3 | -42.5 |
| | 16 | 80.2 | 69.6 | 1.5 | 11.4 | 24.6 | 11.9 | 39.1 | 3.5 | 11.2 | 23.3 | 27.6 | -28.2 |
| *Similarity* | 4 | **94.8** | 68.3 | **60.0** | 30.0 | 73.3 | 60.5 | 47.7 | 55.6 | 47.9 | 66.4 | 60.4 | +4.6 |
| | 8 | 94.1 | 71.0 | 51.9 | 30.6 | 73.1 | 63.7 | 47.2 | **55.7** | 51.2 | 69.3 | **60.8** | +5.0 |
| | 16 | 94.6 | 72.8 | 53.8 | 29.7 | **73.4** | 64.3 | 46.0 | 52.4 | **51.8** | 69.2 | **60.8** | +5.0 |
| *Phi-3.5-Vision* | | | | | | | | | | | | | |
| *Vanilla* | - | 76.4 | 52.4 | 33.7 | 6.8 | 27.3 | 53.1 | 49.4 | 6.5 | 31.8 | 47.6 | 38.5 | |
| *Random* | 4 | 60.5 | 23.6 | 23.6 | 2.0 | 10.2 | 21.1 | 23.4 | 1.3 | 12.9 | 16.3 | 19.5 | -19.0 |
| | 8 | 47.9 | 11.8 | 18.1 | 1.4 | 5.7 | 14.1 | 11.5 | 0.9 | 7.1 | 8.2 | 12.7 | -25.7 |
| | 16 | 55.7 | 26.8 | 14.0 | 2.0 | 32.0 | 6.8 | 42.1 | 0.6 | 2.9 | 21.7 | 20.5 | -18.0 |
| *Similarity* | 4 | 86.5 | 39.5 | 50.7 | 13.6 | 65.3 | 58.8 | 59.2 | 27.4 | 49.2 | 54.7 | 50.5 | +12.0 |
| | 8 | 91.2 | 51.1 | 65.6 | 17.5 | 74.6 | 65.1 | 67.0 | 36.0 | 58.0 | 65.3 | 59.1 | +20.6 |
| | 16 | **94.3** | **64.6** | **73.7** | **23.7** | **83.2** | **73.0** | **76.6** | **47.2** | **67.0** | **73.2** | **67.7** | +29.2 |
| *Phi-4-MM* | | | | | | | | | | | | | |
| *Vanilla* | - | 85.8 | 56.6 | 29.5 | 14.7 | 25.5 | 54.2 | **47.7** | 3.3 | 29.3 | 48.9 | 39.6 | |
| *Random* | 4 | 23.2 | 16.9 | 15.9 | 3.3 | 7.3 | 19.6 | 15.5 | 1.2 | 10.7 | 8.2 | 12.2 | -27.4 |
| | 8 | 22.0 | 10.2 | 13.7 | 1.7 | 10.0 | 11.2 | 12.2 | 1.3 | 7.0 | 8.6 | 9.8 | -29.8 |
| | 16 | 63.2 | 49.3 | 12.9 | 9.4 | 23.1 | 6.8 | 8.6 | 0.9 | 3.9 | 14.9 | 19.3 | -20.3 |
| *Similarity* | 4 | 84.7 | 57.8 | 36.4 | 24.5 | 58.9 | 49.0 | 45.2 | 33.1 | 45.2 | 43.1 | 47.8 | +8.2 |
| | 8 | **86.0** | 68.4 | 42.9 | **29.6** | 64.6 | 54.4 | 42.6 | 32.9 | 49.2 | **45.5** | 51.6 | +12.0 |
| | 16 | 81.9 | **70.6** | **43.2** | 28.8 | **69.5** | **60.5** | 45.1 | **34.2** | **53.9** | 42.3 | **53.0** | +13.4 |

absent when using *Random* or *Pseudo*-labeled contexts.

Table 5. **Open-world results.** We report results for *Llama Inclusion* (LI), *Semantic Similarity* (SS), *Concept Similarity* (bCS), and *Median Concept Similarity* (mCS). Purple indicates our CIRCLE. Higher is better on all metrics. For each LMM, **bold** indicates the best result. Results are split in Tabs. 5a to 5c for readability.

(a) **Qwen2-VL and Qwen2.5-VL.**

| Method | Shots | Prototypical | | | | Non-prototypical | | | | Fine-grained | | | | Very fine-grained | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS |
| *Qwen2-VL 7B* | | | | | | | | | | | | | | | | | |
| *Zero-Shot* | - | 78.7 | 51.9 | 76.0 | 43.7 | 42.6 | 30.8 | 49.8 | 29.2 | 64.0 | 39.2 | 62.9 | 31.9 | 63.0 | 34.5 | 43.4 | 33.1 |
| *Random Ctx* | 4 | 48.0 | 48.9 | 67.7 | 39.5 | 30.0 | 31.2 | 49.4 | 29.0 | 42.9 | 40.5 | 55.9 | 37.3 | 41.0 | 37.0 | 46.3 | 34.7 |
| | 8 | 48.2 | 49.2 | 66.0 | 41.9 | 24.0 | 27.6 | 45.7 | 26.6 | 37.6 | 36.8 | 49.7 | 35.7 | 37.0 | 31.4 | 39.2 | 29.7 |
| | 16 | 24.4 | 41.4 | 52.7 | 39.7 | 17.1 | 23.4 | 41.3 | 23.0 | 31.7 | 34.4 | 44.8 | 34.6 | 31.1 | 29.2 | 34.1 | 27.9 |
| *Pseudo ICL* | 4 | 81.1 | 53.4 | **76.2** | 44.4 | 42.8 | 31.2 | 50.1 | 26.9 | 53.1 | 40.2 | 64.4 | 30.7 | 49.1 | 38.9 | 49.1 | 38.6 |
| | 8 | 76.8 | 52.9 | 75.2 | 44.6 | 37.7 | 34.0 | **52.1** | 31.9 | 47.3 | 36.6 | 58.9 | 32.4 | 47.0 | 37.1 | 46.4 | 36.3 |
| | 16 | 73.7 | 52.6 | 74.0 | 46.3 | 35.1 | 30.3 | 47.9 | 27.3 | 48.1 | 37.9 | 59.4 | 33.3 | 49.8 | 36.9 | 45.9 | 35.7 |
| CIRCLE | 4 | 90.8 | 59.8 | 73.7 | 60.5 | **65.9** | 36.9 | 46.7 | 37.5 | **88.8** | 55.1 | 66.4 | 52.1 | 93.4 | 36.1 | 44.9 | 35.8 |
| | 8 | **92.1** | 61.4 | 72.0 | 59.6 | 62.1 | 37.9 | 45.3 | 36.6 | 86.6 | 58.9 | 69.8 | 54.0 | **95.6** | 35.2 | 43.0 | 35.4 |
| | 16 | 91.5 | **65.6** | 74.3 | **63.5** | 61.6 | **41.9** | 49.4 | **40.5** | 87.3 | **61.1** | **72.0** | **57.3** | 91.5 | **42.5** | **50.2** | **39.8** |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | | | | | |
| *Zero-Shot* | - | 82.9 | 47.9 | **79.9** | 31.1 | 45.9 | 30.5 | 54.0 | 24.8 | 73.8 | 47.0 | **78.9** | 29.5 | 69.0 | 45.8 | 68.6 | 27.1 |
| *Random Ctx* | 4 | 81.9 | 52.3 | 77.8 | 41.6 | 48.5 | 34.5 | 57.5 | 26.6 | 72.3 | 48.5 | 76.9 | 34.4 | 35.6 | 50.3 | 68.4 | 38.1 |
| | 8 | 82.1 | 53.0 | 78.0 | 42.8 | 48.5 | 34.9 | 58.2 | 26.8 | 72.1 | 49.0 | 76.8 | 36.2 | 34.4 | 52.0 | 69.8 | 40.8 |
| | 16 | 82.5 | 53.5 | 78.2 | 43.1 | 48.2 | 35.7 | **58.9** | 27.3 | 70.6 | 49.0 | 76.7 | 36.3 | 34.7 | **52.4** | **70.2** | **42.0** |
| *Pseudo ICL* | 4 | 80.6 | 49.3 | 78.6 | 31.0 | 42.7 | 31.6 | 53.2 | 24.2 | 63.8 | 45.1 | 74.8 | 28.4 | 39.7 | 46.0 | 63.4 | 25.5 |
| | 8 | 80.5 | 49.3 | 78.8 | 30.6 | 43.7 | 31.9 | 53.3 | 24.4 | 63.6 | 44.9 | 75.2 | 27.9 | 40.5 | 46.3 | 64.9 | 24.0 |
| | 16 | 80.7 | 49.3 | 79.0 | 30.9 | 42.9 | 32.3 | 53.7 | 24.5 | 65.5 | 45.7 | 75.5 | 28.5 | 35.9 | 47.6 | 65.5 | 25.0 |
| CIRCLE | 4 | 90.3 | 68.0 | 69.7 | 67.0 | 61.0 | 40.1 | 44.1 | 39.5 | 82.3 | 60.4 | 63.2 | 59.2 | 88.7 | 39.6 | 39.6 | 39.6 |
| | 8 | 89.7 | **70.0** | 71.2 | **69.4** | 58.5 | 39.0 | 39.6 | 39.1 | 80.5 | **61.4** | 62.6 | **60.9** | 87.6 | 37.9 | 37.9 | 37.9 |
| | 16 | **94.9** | 67.7 | 68.1 | 67.2 | **67.6** | **42.6** | 45.1 | **42.3** | **86.3** | 60.1 | 60.9 | 59.7 | **93.6** | 36.4 | 36.6 | 36.5 |

(b) **LLaVa OneVision**.

| Method | Shots | Prototypical | | | | Non-prototypical | | | | Fine-grained | | | | Very fine-grained | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS |
| *LLaVa OneVision 7B* | | | | | | | | | | | | | | | | | |
| *Zero-Shot* | - | 53.2 | 56.2 | 62.0 | 53.4 | 28.1 | 31.6 | 43.8 | 30.2 | 40.4 | 39.0 | 43.9 | 37.2 | **76.7** | 31.8 | 32.3 | 30.9 |
| *Random Ctx* | 4 | 13.9 | 33.7 | 35.1 | 33.9 | 7.3 | 25.3 | 35.7 | 23.7 | 20.8 | 34.2 | 34.5 | 34.2 | 74.1 | 30.8 | 30.8 | 30.8 |
| | 8 | 14.2 | 33.1 | 35.5 | 33.2 | 6.8 | 25.9 | 37.9 | 24.0 | 20.7 | 34.3 | 34.6 | 34.3 | 74.1 | 30.8 | 30.8 | 30.8 |
| | 16 | 14.0 | 29.3 | 36.7 | 29.6 | 8.6 | 26.0 | 39.3 | 24.1 | 21.0 | 33.2 | 35.8 | 33.4 | 75.8 | 30.8 | 30.8 | 30.8 |
| *Pseudo ICL* | 4 | 19.7 | 31.2 | 41.1 | 30.5 | 3.3 | 18.3 | 31.5 | 19.7 | 20.3 | 35.5 | 40.0 | 34.7 | 70.4 | 30.6 | 31.1 | 29.8 |
| | 8 | 55.4 | 54.1 | 63.9 | 48.9 | 29.9 | 27.5 | 45.6 | 24.0 | 33.9 | 37.1 | 45.7 | 34.2 | 74.4 | 31.3 | 31.6 | 30.7 |
| | 16 | 58.1 | 55.0 | 65.3 | 49.3 | 31.8 | 28.1 | 46.3 | 25.1 | 33.6 | 37.1 | 45.7 | 34.5 | 73.3 | 31.7 | 32.3 | 30.5 |
| CIRCLE | 4 | 84.9 | 71.8 | 71.8 | 71.8 | 52.5 | 41.7 | 41.8 | 41.8 | 70.2 | 42.7 | 42.7 | 42.6 | 67.4 | **35.1** | **35.3** | **34.6** |
| | 8 | **87.4** | 73.8 | 73.8 | 73.8 | 55.3 | 46.3 | 46.3 | 46.3 | **73.9** | 41.3 | 41.3 | 41.3 | 54.9 | 32.3 | 32.3 | 32.3 |
| | 16 | 72.2 | **74.0** | **74.0** | **74.0** | **61.7** | **55.3** | **55.3** | **55.3** | 55.1 | **46.0** | **46.0** | **46.0** | 74.2 | 32.9 | 32.8 | 32.9 |

## D. Streaming results

We provide detailed results for the Open-World *Streaming* experiments in this section. We evaluate all five models on the ten datasets, testing the *Pseudo-label* ICL variant and CIRCLE.

We organize the results as follows:

- **Main results.** Tab. 10 presents the aggregated performance for the Qwen series, LLaVa OneVision, and the Phi series, categorizing results by dataset group: *Prototypical*, *Non-prototypical*, *Fine-grained*, and *Very fine-grained*.
- **Metric-specific results.** Tabs. 11 to 14 provide granular breakdowns across the ten datasets for SS, bCS, mCS, and LI, respectively. We report the average score and the improvement ($\Delta$) relative to the *Zero-Shot* baseline.

**Results discussion.** The streaming results in Tab. 10 confirm the robustness of CIRCLE in a streaming scenario. Across the majority of models and dataset groups, CIRCLE con-

(c) **Phi-3.5-Vision and Phi-4-Multimodal.**

| Method | Shots | Prototypical | | | | Non-prototypical | | | | Fine-grained | | | | Very fine-grained | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS | LI | SS | bCS | mCS |
| *Phi-3.5-Vision* | | | | | | | | | | | | | | | | | |
| *Zero-Shot* | - | 60.7 | 48.2 | 65.6 | 46.1 | 28.7 | 24.9 | 36.7 | 24.1 | 50.7 | 32.1 | 47.2 | 31.3 | 54.2 | 29.5 | 36.3 | 29.8 |
| *Random Ctx* | 4 | 49.2 | 57.5 | 60.9 | 57.3 | 14.7 | 27.2 | 35.4 | 27.2 | 29.1 | 35.9 | 43.1 | 35.6 | 51.7 | 27.7 | 32.3 | 26.5 |
| | 8 | 45.0 | 53.4 | 57.6 | 53.3 | 11.3 | 27.1 | 37.5 | 27.5 | 26.7 | 34.7 | 41.5 | 34.7 | 56.7 | 28.2 | 31.7 | 27.3 |
| | 16 | 40.3 | 48.3 | 54.4 | 48.4 | 10.4 | 26.3 | 36.4 | 27.4 | 24.1 | 26.3 | 36.4 | 27.4 | 58.2 | 27.1 | 31.7 | 26.3 |
| *Pseudo ICL* | 4 | 54.1 | 44.1 | 61.7 | 40.1 | 23.7 | 22.8 | 35.1 | 21.5 | 43.1 | 33.0 | 48.6 | 29.7 | 24.9 | 32.4 | **41.8** | 32.5 |
| | 8 | 59.3 | 46.7 | 65.3 | 41.1 | 24.1 | 22.3 | 34.6 | 21.3 | 34.5 | 32.3 | 50.0 | 28.2 | 25.8 | 32.8 | 41.0 | 34.0 |
| | 16 | 55.7 | 46.1 | 63.7 | 43.1 | 25.2 | 22.9 | 34.9 | 21.4 | 35.0 | 33.8 | **51.9** | 29.5 | 30.0 | 33.3 | 40.0 | 33.6 |
| CIRCLE | 4 | 80.8 | 63.3 | 64.6 | 63.1 | 52.3 | **36.5** | **37.6** | **36.7** | 84.8 | 39.1 | 44.9 | 39.2 | 88.8 | **36.6** | 38.1 | **36.1** |
| | 8 | 84.7 | **64.7** | **66.2** | **65.0** | **59.3** | 31.6 | 36.0 | 32.9 | 84.7 | **42.1** | 48.9 | **42.6** | 92.9 | 32.0 | 37.3 | 32.4 |
| | 16 | **92.1** | 59.7 | 63.2 | 60.3 | 58.3 | 30.0 | 35.2 | 31.7 | **88.1** | 39.2 | 45.7 | 42.1 | **99.6** | 33.0 | 33.6 | 33.1 |
| *Phi-4-Multimodal* | | | | | | | | | | | | | | | | | |
| *Zero-Shot* | - | 49.8 | 57.4 | 58.7 | 57.2 | 21.2 | 29.2 | 32.7 | 29.2 | 37.7 | 39.2 | 39.2 | 39.1 | 73.6 | 31.6 | 31.7 | 31.6 |
| *Random Ctx* | 4 | 10.9 | 32.6 | 32.6 | 32.6 | 9.6 | 30.1 | 30.2 | 30.1 | 29.9 | 38.1 | 38.1 | 38.1 | 70.0 | 30.6 | 30.6 | 30.6 |
| | 8 | 13.0 | 34.2 | 34.2 | 34.2 | 7.7 | 29.6 | 29.7 | 29.6 | 31.2 | 38.5 | 38.5 | 38.5 | 71.7 | 30.7 | 30.7 | 30.7 |
| | 16 | 11.3 | 32.8 | 32.8 | 32.8 | 5.8 | 28.5 | 28.6 | 28.5 | 30.9 | 37.9 | 37.9 | 37.9 | 72.3 | 30.8 | 30.7 | 30.6 |
| *Pseudo ICL* | 4 | 51.9 | 61.5 | 62.0 | 61.5 | 15.1 | 26.6 | 31.2 | 26.7 | 37.7 | 41.6 | 41.7 | 41.5 | 72.8 | 31.9 | 31.9 | 31.9 |
| | 8 | 49.2 | 59.9 | 60.0 | 59.9 | 14.9 | 32.0 | 33.3 | 32.0 | 36.4 | 41.4 | 41.4 | 41.4 | 73.2 | 31.5 | 31.5 | 31.5 |
| | 16 | 51.6 | 61.5 | 61.5 | 61.5 | 19.1 | 24.7 | 31.1 | 24.5 | 35.7 | 41.1 | 41.1 | 41.0 | 71.6 | 32.0 | 32.0 | 32.0 |
| CIRCLE | 4 | 87.0 | **68.5** | 69.4 | **68.6** | 40.0 | 35.9 | 36.6 | 35.6 | **80.7** | 49.7 | 53.2 | 49.2 | **92.4** | 34.6 | 36.3 | 34.4 |
| | 8 | 85.3 | 65.9 | 68.4 | 65.7 | 61.5 | 37.4 | 43.8 | 34.7 | 79.3 | **57.0** | **60.3** | **56.8** | 91.5 | 38.2 | 38.9 | 37.7 |
| | 16 | **91.5** | 65.5 | **70.1** | 66.4 | **67.6** | **43.2** | **46.1** | **43.4** | 79.1 | 53.3 | 55.5 | 53.0 | 75.2 | **40.2** | **42.5** | **37.9** |

sistently outperforms both the *Zero-Shot* baseline and the *Pseudo ICL* approach. For example, on *Prototypical* datasets, CIRCLE achieves a LI of 90.4 with Qwen2-VL (vs. 78.7 *Zero-Shot*) and 84.9 with Phi-4-Multimodal (vs. 49.8 *Zero-Shot*). Similarly, SS increases from 51.9 to 60.9 for Qwen2-VL and from 57.4 to 62.7 for Phi-4-Multimodal, while mCS from 43.7 to 59.5 and from 57.2 to 64.6, respectively. While *Pseudo ICL* occasionally shows strength in specific similarity metrics (*e.g.*, SS for Phi-3.5-Vision), it is inconsistent across models, metrics, and datasets. In contrast, CIRCLE exhibits superior overall stability.

## E. Implementation details

For the closed-world experiments, we evaluate the three most common CLIP [30] variants using ViT-B/32, ViT-B/16, and ViT-L/14 [13]. For both the closed-world and open-world experiments, we evaluate five Large Multimodal Models (LMMs) from three publicly available model series: (i) Qwen2-VL 7B [37] and Qwen2.5-VL 7B [3]; (ii) LLaVa OneVision 7B [20]; (iii) Phi-3.5-Vision [1] and Phi-4-Multimodal [25].

For LMMs, we use a varying batch size for generation. For tasks with a small context (*e.g.*, multi-choice experiments), our hardware (see *Resources* below) supports up to batch size 32 for Qwen2-VL and Qwen2.5-VL. For more complex and rich contexts, such as that of few-shot classification in the closed-world setting (see Sec. 3 and Tab. 1) and the context we build for CIRCLE (see Sec. 4 and Tab. 2), the batch size needs to be decreased to 8, 4, or 2 samples per GPU, depending on the LMM. To reduce VRAM usage, we downscale the context images to $224 \times 224$ pixels. For reproducibility, we always use greedy decoding up to 64 generated tokens.

For a fair comparison, we use the same experimental framework of [12] for all our experiments. We note that the Llama Inclusion metric is sensitive, and provides incorrect scores for non-sentence responses. Therefore, we encapsulate the comma-separated class options provided by CIRCLE in the template "The target object in the photo is one of these [output].".

We run all our experiments on NVIDIA A100 GPUs with 40, 64, and 80 GB of VRAM. Simpler experiments (*e.g.*, *Zero-shot* closed-world) can be run on a single GPU, while we run experiments with large contexts on multiple GPUs to reduce wait times, using up to 4 GPUs per experiment. Evaluation time ranges from a few minutes for *Zero-shot* closed-world experiments to 8-10 hours for the largest datasets (Food101 and SUN397, see Tab. 3 for details on their sizes) in the *Streaming* setting, where the context has to be re-generated many times.

## F. Qualitative results

We provide additional qualitative examples to complement the visual analysis presented in Fig. 4 (*Main*). Figs. 7 to 13

Table 6. **Open-world results.** ***Semantic Similarity*** (SS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. Δ computed w.r.t. the average (Avg.) of *Zero-Shot*. Results are split in Tabs. 6a to 6c for readability.

(a) **Qwen2-VL and Qwen2.5-VL.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 55.8 | 28.6 | 20.7 | 20.6 | 41.7 | 50.7 | 25.1 | 48.3 | 48.1 | 43.1 | 38.3 | |
| *Random Ctx* | 4 | 54.8 | 25.5 | 27.3 | 20.8 | 35.7 | 52.5 | 33.1 | 53.3 | 42.9 | 41.0 | 38.7 | -0.4 |
| | 8 | 56.1 | 21.5 | 27.9 | 20.3 | 30.8 | 50.9 | 28.7 | 42.5 | 42.3 | 33.3 | 35.4 | -2.9 |
| | 16 | 50.8 | 18.8 | 26.4 | 22.8 | 26.7 | 46.5 | 30.0 | 35.5 | 32.0 | 25.1 | 31.5 | -6.8 |
| *Pseudo ICL* | 4 | 57.6 | 29.7 | 18.6 | 21.0 | 40.5 | 52.6 | 27.3 | **56.8** | 49.2 | 45.5 | 39.9 | +1.6 |
| | 8 | 57.1 | 28.4 | 29.1 | 20.5 | 35.6 | 50.1 | 24.1 | 53.6 | 48.6 | 44.5 | 39.2 | +0.9 |
| | 16 | 58.2 | 29.3 | 16.6 | 20.3 | 39.0 | 49.4 | 25.4 | 53.5 | 46.9 | 44.9 | 38.4 | +0.1 |
| CIRCLE | 4 | 62.5 | 29.0 | 32.2 | 26.3 | 49.3 | 66.1 | 49.7 | 45.9 | 57.0 | 49.6 | 46.8 | +8.5 |
| | 8 | 61.3 | 35.7 | 31.3 | 25.0 | **57.3** | 66.4 | 53.1 | 45.4 | 61.5 | 46.9 | 48.4 | +10.5 |
| | 16 | **68.4** | **36.5** | **37.0** | **31.4** | 54.5 | **67.9** | **61.0** | 53.7 | **62.7** | **52.3** | **52.5** | +14.2 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 48.8 | 28.3 | 19.0 | 36.7 | 47.4 | 52.4 | 41.1 | 54.9 | 47.0 | 44.2 | 42.0 | |
| *Random Ctx* | 4 | 55.8 | 29.1 | 28.3 | 35.4 | 49.6 | 47.7 | 48.1 | 65.2 | 48.8 | 46.2 | 45.4 | +3.4 |
| | 8 | 56.7 | 29.9 | 29.5 | **36.0** | 50.1 | 47.7 | 49.2 | 68.1 | 49.3 | 45.3 | 46.2 | +4.2 |
| | 16 | 57.6 | 30.2 | 31.7 | 35.8 | 51.0 | 47.8 | 48.2 | **69.0** | 49.5 | 45.1 | 46.6 | +4.6 |
| *Pseudo ICL* | 4 | 50.8 | 29.5 | 19.1 | 33.0 | 47.5 | 49.0 | 38.9 | 58.9 | 47.9 | 46.0 | 42.1 | +0.1 |
| | 8 | 50.3 | 29.8 | 19.2 | 34.0 | 48.4 | 49.2 | 37.1 | 58.6 | 48.3 | 46.6 | 42.2 | +0.2 |
| | 16 | 50.2 | 30.4 | 20.6 | 34.7 | 47.3 | 50.1 | 39.5 | 60.6 | 48.5 | 46.1 | 42.8 | +0.8 |
| CIRCLE | 4 | 68.0 | 39.8 | 30.9 | 29.6 | 64.5 | **58.4** | 58.4 | 49.5 | **68.0** | 49.5 | **51.7** | +9.7 |
| | 8 | **73.6** | **42.3** | 34.4 | 29.0 | **66.9** | **58.4** | **58.9** | 46.7 | 66.3 | 40.4 | **51.7** | +9.7 |
| | 16 | 70.3 | 41.7 | **35.1** | 29.6 | 66.4 | 56.8 | 57.1 | 43.3 | 65.0 | **51.0** | 51.6 | +9.6 |

(b) **LLaVa OneVision**.

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *LLaVa OneVision 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 68.9 | 32.0 | 19.4 | 29.4 | 37.5 | 41.6 | 37.8 | 34.3 | 43.4 | 43.4 | 38.8 | |
| *Random Ctx* | 4 | 41.2 | 24.7 | 24.9 | 29.4 | 32.1 | 33.0 | 37.6 | 32.2 | 26.3 | 26.3 | 30.8 | -8.0 |
| | 8 | 40.1 | 24.1 | 27.1 | 29.4 | 32.0 | 33.3 | 37.7 | 32.2 | 26.1 | 26.4 | 30.8 | -8.0 |
| | 16 | 31.5 | 25.3 | 26.3 | 29.4 | 27.8 | 34.0 | 37.8 | 32.2 | 27.2 | 26.3 | 29.8 | -9.0 |
| *Pseudo ICL* | 4 | 19.9 | 17.2 | 17.6 | 29.4 | 26.3 | 45.1 | 35.1 | 31.8 | 42.5 | 20.0 | 28.5 | -10.3 |
| | 8 | 67.0 | 31.0 | 23.5 | 29.4 | 33.8 | 39.5 | 38.1 | 33.2 | 41.2 | 27.9 | 36.4 | -2.4 |
| | 16 | 66.5 | 30.1 | 23.2 | 29.4 | 32.7 | 40.6 | 38.2 | 34.0 | 43.5 | 31.0 | 36.9 | -1.9 |
| CIRCLE | 4 | 79.4 | 61.7 | 23.8 | 29.4 | **48.7** | 41.0 | 38.4 | **40.8** | 64.2 | 39.7 | 46.7 | +7.9 |
| | 8 | 81.1 | 52.2 | 32.9 | **29.8** | 40.6 | 43.9 | 39.3 | 34.8 | **66.5** | 53.9 | 47.5 | +8.7 |
| | 16 | **82.7** | **71.8** | **36.7** | **29.8** | 41.7 | **53.8** | **42.5** | 35.9 | 65.3 | **57.4** | **51.8** | +13.0 |

display a selection of samples across a subset of the ten evaluated datasets, providing three distinct examples for each to illustrate the model's behavior in diverse scenarios.

In these examples, we observe that while baseline methods exhibit inconsistent performance, occasionally identifying the correct concept but frequently misclassifying or hallucinating unrelated categories (especially in the *Random* context), our CIRCLE predicts the correct label more consistently. This underscores the robustness of CIRCLE, which effectively leverages the context it constructs during the iterative refinement steps to guide the underlying LMM to the correct specificity and format.

(c) **Phi-3.5-Vision and Phi-4-Multimodal.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phi-3.5-Vision* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 53.2 | 29.1 | 7.4 | 19.9 | 31.6 | 40.2 | 24.6 | 39.1 | 43.2 | 38.3 | 32.6 | |
| *Random Ctx* | 4 | 67.0 | 26.9 | 16.9 | 20.5 | 36.3 | 41.8 | 29.6 | 34.9 | 48.1 | 37.8 | 36.0 | +3.4 |
| | 8 | 62.1 | 24.7 | 24.1 | 22.9 | 36.6 | 37.5 | 30.1 | 33.5 | 44.8 | 32.6 | 34.9 | +2.3 |
| | 16 | 57.4 | 24.2 | 25.1 | 21.0 | 35.3 | 34.2 | 29.4 | 33.2 | 39.3 | 29.7 | 32.9 | +2.3 |
| *Pseudo ICL* | 4 | 49.6 | 26.0 | 6.6 | 18.8 | 31.7 | 39.3 | 28.0 | 46.0 | 38.5 | 35.8 | 32.1 | -0.5 |
| | 8 | 50.6 | 27.3 | 5.5 | 19.0 | 32.8 | 40.3 | 23.7 | 46.6 | 42.9 | 34.0 | 32.3 | -0.3 |
| | 16 | 52.7 | 26.4 | 5.8 | 19.8 | 31.0 | 40.6 | 29.9 | **46.8** | 39.6 | 36.4 | 32.9 | +0.3 |
| CIRCLE | 4 | **69.7** | 37.8 | **32.0** | 28.8 | **45.2** | 38.2 | 34.0 | 44.4 | 56.8 | **39.6** | 42.7 | +10.1 |
| | 8 | 72.4 | **37.9** | 30.0 | 25.5 | 42.8 | **49.0** | **34.4** | 38.5 | **56.9** | 26.9 | 41.4 | +8.8 |
| | 16 | 67.9 | 32.6 | 31.0 | **29.8** | 42.2 | 44.0 | 31.5 | 36.3 | 51.4 | 26.6 | **39.3** | +6.7 |
| *Phi-4-Multimodal* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 73.5 | 33.9 | 13.8 | 29.2 | 42.2 | 37.8 | 37.5 | 34.1 | 41.2 | 40.0 | 38.3 | |
| *Random Ctx* | 4 | 36.7 | 29.7 | 36.1 | 28.8 | 40.9 | 36.5 | 36.8 | 32.3 | 28.5 | 24.4 | 33.1 | -5.2 |
| | 8 | 39.5 | 27.8 | 36.9 | 29.1 | 41.4 | 36.4 | 37.6 | 32.3 | 28.9 | 24.2 | 33.4 | -4.9 |
| | 16 | 38.2 | 26.4 | 36.0 | 29.3 | 41.0 | 35.2 | 37.6 | 32.2 | 27.4 | 23.2 | 32.6 | -5.7 |
| *Pseudo ICL* | 4 | 75.7 | 32.2 | 14.9 | 29.3 | 42.8 | 43.8 | 38.2 | 34.5 | 47.4 | 32.7 | 39.1 | +0.8 |
| | 8 | 74.4 | 36.9 | 25.1 | **29.4** | 42.0 | 43.9 | 38.2 | 33.6 | 45.4 | 34.0 | 40.3 | +2.0 |
| | 16 | 76.4 | 29.4 | 13.2 | **29.4** | 41.9 | 43.1 | 38.2 | 34.7 | 46.7 | 31.5 | 38.4 | +0.1 |
| CIRCLE | 4 | **77.2** | 28.8 | 27.5 | 28.7 | 49.2 | 53.3 | 46.7 | 40.5 | 59.7 | **51.5** | 46.3 | +8.2 |
| | 8 | 71.9 | 38.3 | 36.8 | 28.4 | **54.6** | **62.1** | 54.3 | 48.1 | **59.9** | 37.0 | 49.1 | +10.8 |
| | 16 | 72.8 | **39.6** | **42.1** | 28.4 | 47.8 | 57.3 | **54.9** | **51.9** | 58.3 | 47.9 | **50.1** | +11.8 |

Table 7. **Open-world results.** ***Concept Similarity*** (bCS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. For each LMM, **bold** indicates the best result. Δ computed w.r.t. the average (Avg.) of *Zero-Shot*. Results are split in Tabs. 7a to 7c for readability.

(a) **Qwen2-VL and Qwen2.5-VL.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | **81.3** | 50.3 | 39.8 | 30.7 | **68.7** | 77.0 | 43.2 | 55.7 | 70.7 | 59.0 | 57.6 | |
| *Random Ctx* | 4 | 75.9 | 42.7 | 52.0 | 30.7 | 49.4 | 68.7 | 49.7 | 61.9 | 59.4 | 53.5 | 54.4 | -3.2 |
| | 8 | 75.4 | 38.3 | 53.7 | 29.6 | 46.1 | 64.2 | 38.7 | 48.8 | 56.5 | 45.2 | 49.7 | -7.9 |
| | 16 | 63.3 | 33.7 | **54.2** | 29.4 | 41.0 | 57.6 | 35.9 | 38.8 | 42.0 | 35.9 | 43.2 | -14.4 |
| *Pseudo ICL* | 4 | 79.9 | **51.1** | 37.2 | 31.3 | 63.7 | 78.4 | 51.0 | **67.0** | **72.4** | **62.1** | 59.4 | +1.8 |
| | 8 | 80.2 | 44.9 | 53.2 | 30.6 | 56.8 | 73.2 | 46.8 | 62.1 | 70.1 | 58.3 | 57.6 | +0.0 |
| | 16 | 80.2 | 48.5 | 34.0 | 30.2 | 58.4 | 72.5 | 47.4 | 61.6 | 67.7 | 61.1 | 56.2 | -1.4 |
| CIRCLE | 4 | 79.6 | 42.6 | 35.3 | 38.1 | 54.3 | **80.4** | 64.7 | 51.8 | 67.8 | **62.1** | 57.7 | +0.1 |
| | 8 | 76.5 | 44.8 | 37.4 | 35.6 | 64.8 | 74.6 | 69.9 | 50.4 | 67.6 | 53.8 | 57.5 | -0.1 |
| | 16 | 79.0 | 46.2 | 40.2 | **42.2** | 64.4 | 77.5 | **73.9** | 58.1 | 69.5 | 61.8 | **61.3** | +3.7 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | **85.6** | 53.4 | 41.3 | **68.7** | **79.7** | **79.6** | 77.3 | 68.5 | **74.2** | **67.2** | 69.5 | |
| *Random Ctx* | 4 | 82.9 | 52.6 | 54.8 | 60.0 | 77.0 | 72.5 | 81.3 | 76.9 | 72.6 | 65.2 | 69.6 | +0.1 |
| | 8 | 82.9 | 52.7 | 56.7 | 60.6 | 76.4 | 72.6 | **81.4** | 79.0 | 73.0 | 65.1 | 70.0 | +0.5 |
| | 16 | 83.5 | 52.9 | **59.0** | 60.3 | 76.6 | 72.6 | 80.8 | **80.1** | 73.0 | 64.7 | **70.3** | +0.8 |
| *Pseudo ICL* | 4 | 85.2 | 53.3 | 41.3 | 57.1 | 77.2 | 72.9 | 74.3 | 69.6 | 72.1 | 65.1 | 66.8 | -2.7 |
| | 8 | 85.0 | 53.3 | 41.0 | 60.0 | 78.6 | 73.4 | 73.7 | 69.9 | 72.7 | 65.4 | 67.3 | -2.2 |
| | 16 | 85.0 | **54.1** | 42.1 | 59.9 | 77.1 | 74.6 | 74.7 | 71.1 | 73.0 | 65.0 | 67.7 | -1.8 |
| CIRCLE | 4 | 71.1 | 47.6 | 30.9 | 29.6 | 64.5 | 60.6 | 64.5 | 49.5 | 68.2 | 53.8 | 54.0 | -15.5 |
| | 8 | 75.7 | 43.7 | 34.4 | 29.1 | 67.7 | 60.3 | 60.0 | 46.7 | 66.7 | 40.6 | 52.5 | -17.0 |
| | 16 | 71.1 | 45.3 | 35.6 | 29.9 | 66.6 | 57.8 | 58.5 | 43.3 | 65.1 | 54.4 | 52.8 | -16.7 |

(b) **LLaVa OneVision.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *LLaVa OneVision 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 79.1 | 46.9 | 41.0 | 29.4 | 51.9 | 41.9 | 37.9 | 35.3 | 44.8 | 43.5 | 45.2 | |
| *Random Ctx* | 4 | 44.0 | 30.3 | 50.4 | 29.4 | 32.7 | 33.0 | 37.6 | 32.2 | 26.3 | 26.3 | 34.2 | -11.0 |
| | 8 | 44.9 | 30.9 | **56.4** | 29.4 | 32.7 | 33.3 | 37.7 | 32.2 | 26.1 | 26.4 | 35.0 | -10.2 |
| | 16 | 46.2 | 36.4 | 55.2 | 29.4 | 35.5 | 34.0 | 37.8 | 32.2 | 27.2 | 26.3 | 36.0 | -9.2 |
| *Pseudo ICL* | 4 | 33.0 | 35.0 | 38.6 | 29.4 | 35.1 | 49.7 | 35.3 | 32.9 | 49.2 | 20.9 | 35.9 | -9.3 |
| | 8 | **84.3** | 53.7 | 54.6 | 29.4 | **58.2** | 40.5 | 38.4 | 33.8 | 43.5 | 28.5 | 46.5 | +1.3 |
| | 16 | 84.1 | 53.5 | 53.9 | 29.4 | 57.0 | 41.6 | 38.4 | 35.2 | 46.6 | 31.5 | 47.1 | +1.9 |
| CIRCLE | 4 | 79.4 | 61.7 | 23.8 | 29.4 | 48.7 | 41.0 | 38.4 | **41.3** | 64.2 | 39.8 | 46.8 | +1.6 |
| | 8 | 81.1 | 52.3 | 32.9 | **29.8** | 40.6 | 43.9 | 39.3 | 34.8 | **66.5** | 53.9 | 47.5 | +2.3 |
| | 16 | 82.7 | **71.8** | 36.7 | **29.8** | 41.7 | **53.9** | **42.5** | 35.9 | 65.3 | **57.4** | **51.8** | +6.6 |

(c) **Phi-3.5-Vision and Phi-4-Multimodal.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phi-3.5-Vision* | | | | | | | | | | | | | |
| *Zero-Shot* | - | **73.6** | **43.5** | 16.4 | 29.6 | 44.9 | 58.1 | 38.7 | 43.0 | 57.6 | **50.1** | 45.5 | |
| *Random Ctx* | 4 | 71.1 | 34.1 | 29.2 | 29.3 | 42.5 | 50.2 | 36.7 | 35.3 | 50.6 | 42.7 | 42.2 | -3.3 |
| | 8 | 67.4 | 32.0 | **41.6** | 29.3 | 41.2 | 46.3 | 36.9 | 34.1 | 47.7 | 38.9 | 41.5 | -4.0 |
| | 16 | 64.7 | 31.9 | 40.1 | 29.3 | 41.0 | 42.7 | 36.6 | 34.1 | 44.2 | 37.2 | 40.2 | -5.3 |
| *Pseudo ICL* | 4 | 69.3 | 40.9 | 16.1 | 29.2 | 41.9 | 58.9 | 45.1 | **54.4** | 54.2 | 48.3 | 45.8 | +0.3 |
| | 8 | 70.2 | 41.7 | 14.6 | 28.2 | 45.3 | 59.4 | 45.3 | 53.8 | **60.4** | 47.4 | 46.6 | +1.1 |
| | 16 | 71.5 | 40.8 | 14.9 | 25.8 | 42.9 | **60.4** | **52.5** | 54.2 | 55.9 | 49.1 | **46.8** | +1.3 |
| CIRCLE | 4 | 71.2 | 40.2 | 32.0 | **31.4** | **53.8** | 40.5 | 40.3 | 44.8 | 58.0 | 40.5 | 45.3 | -0.2 |
| | 8 | **73.6** | 42.6 | 30.2 | 30.8 | 52.0 | 52.6 | 42.1 | 43.8 | 58.8 | 35.3 | 46.2 | +0.7 |
| | 16 | 71.0 | 39.0 | 32.6 | 30.9 | 50.5 | 48.2 | 38.4 | 36.3 | 55.3 | 34.1 | 43.6 | -1.9 |
| *Phi-4-Multimodal* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 75.8 | 40.6 | 17.4 | **29.3** | 42.2 | 37.9 | 37.5 | 34.1 | 41.7 | 40.1 | 39.6 | |
| *Random Ctx* | 4 | 36.8 | 29.7 | 36.3 | 28.8 | 40.9 | 36.5 | 36.8 | 32.3 | 28.5 | 24.8 | 33.1 | -6.5 |
| | 8 | 39.5 | 27.8 | 37.2 | 29.1 | 41.4 | 36.4 | 37.6 | 32.3 | 28.9 | 24.2 | 33.4 | -6.2 |
| | 16 | 38.2 | 26.4 | 36.2 | 29.2 | 41.0 | 35.2 | 37.6 | 32.2 | 27.4 | 23.2 | 32.7 | -6.9 |
| *Pseudo ICL* | 4 | 76.7 | 38.5 | 22.4 | **29.3** | 42.9 | 43.9 | 38.2 | 34.5 | 47.4 | 32.7 | 40.6 | +1.0 |
| | 8 | 74.5 | 37.0 | 28.9 | **29.3** | 42.0 | 43.9 | 38.2 | 33.6 | 45.4 | 34.0 | 40.7 | +1.1 |
| | 16 | 76.4 | 41.0 | 21.0 | **29.3** | 41.9 | 43.1 | 38.2 | 34.7 | 46.7 | 31.5 | 40.4 | +0.8 |
| CIRCLE | 4 | 77.8 | 29.3 | 27.5 | **29.3** | 52.7 | 59.0 | 47.9 | 43.3 | **61.1** | 53.1 | 48.1 | +8.5 |
| | 8 | 76.7 | 42.0 | 37.7 | 28.4 | **57.6** | **65.3** | **57.9** | 49.4 | 60.2 | 51.9 | 52.7 | +13.1 |
| | 16 | **81.0** | **42.3** | **42.2** | 29.1 | 50.3 | 59.4 | 56.8 | **55.9** | 59.2 | **53.7** | **53.0** | +13.4 |

Table 8. **Open-world results.** ***Median Concept Similarity*** (mCS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. For each LMM, **bold** indicates the best result. Δ computed w.r.t. the average (Avg.) of *Zero-Shot*. Results are split in Tabs. 8a to 8c for readability.

(a) **Qwen2-VL and Qwen2.5-VL.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 51.3 | 27.9 | 20.6 | 20.4 | 33.3 | 37.3 | 25.0 | 45.8 | 36.1 | 38.9 | 33.7 | |
| *Random Ctx* | 4 | 46.3 | 26.0 | 24.0 | 20.8 | 34.9 | 45.7 | 31.3 | 48.6 | 32.7 | 37.1 | 34.7 | +1.0 |
| | 8 | 49.0 | 23.5 | 23.3 | 20.3 | 30.7 | 47.6 | 28.9 | 39.0 | 34.8 | 33.0 | 33.0 | -0.7 |
| | 16 | 48.6 | 20.8 | 23.1 | 22.8 | 26.8 | 46.8 | 30.0 | 32.9 | 30.8 | 25.2 | 30.8 | -2.9 |
| *Pseudo ICL* | 4 | 56.1 | 27.9 | 18.3 | 20.7 | 30.1 | 35.0 | 27.1 | 56.5 | 32.8 | 34.4 | 33.9 | +0.2 |
| | 8 | 55.2 | 27.7 | 26.3 | 20.3 | 32.2 | 39.9 | 25.1 | 52.3 | 34.1 | 41.9 | 35.5 | +1.8 |
| | 16 | 57.5 | 28.2 | 16.0 | 20.1 | 34.8 | 38.7 | 26.5 | 51.3 | 35.2 | 37.7 | 34.6 | +0.9 |
| CIRCLE | 4 | **66.6** | 29.0 | 34.6 | 26.2 | 52.2 | 62.9 | 41.3 | 45.5 | 54.5 | 48.8 | 46.1 | +12.4 |
| | 8 | 58.3 | 34.3 | 27.4 | 25.7 | **57.3** | 62.9 | 41.8 | 45.1 | 61.0 | 48.2 | 46.2 | +12.5 |
| | 16 | 65.2 | **34.8** | **35.9** | **30.2** | 53.5 | **67.2** | **51.2** | **49.4** | **61.8** | **50.9** | **50.0** | +16.3 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 33.0 | 26.3 | 18.6 | 24.7 | 29.5 | 31.8 | 27.3 | 29.4 | 29.3 | 29.6 | 27.9 | |
| *Random Ctx* | 4 | 51.4 | 26.7 | 21.4 | 26.1 | 33.3 | 31.1 | 38.9 | 50.1 | 31.9 | 31.6 | 34.3 | +6.4 |
| | 8 | 53.1 | 26.9 | 21.9 | 27.2 | 37.4 | 31.0 | 40.1 | 54.3 | 32.4 | 31.6 | 35.6 | +7.7 |
| | 16 | 54.0 | 26.8 | 23.2 | 28.6 | 39.3 | 31.2 | 38.5 | 55.4 | 32.2 | 32.1 | 36.1 | +8.2 |
| *Pseudo ICL* | 4 | 32.6 | 26.7 | 18.9 | 24.3 | 29.4 | 31.1 | 24.8 | 26.8 | 29.5 | 26.9 | 27.1 | -0.8 |
| | 8 | 32.1 | 26.7 | 18.9 | 24.1 | 29.1 | 30.6 | 23.8 | 23.9 | 29.1 | 27.7 | 26.6 | -1.3 |
| | 16 | 32.5 | 27.1 | 20.3 | 24.7 | 28.8 | 31.1 | 25.6 | 25.4 | 29.3 | 26.2 | 27.1 | -0.8 |
| CIRCLE | 4 | 66.0 | 37.9 | 30.9 | **29.6** | 64.2 | **57.9** | 55.6 | **49.5** | **67.9** | 49.9 | 50.9 | +23.0 |
| | 8 | **72.4** | **42.5** | **34.4** | 29.0 | 66.4 | 57.5 | **58.9** | 46.7 | 66.3 | 40.4 | **51.5** | +23.6 |
| | 16 | 69.5 | 41.8 | **34.4** | **29.6** | **66.0** | 56.3 | 56.8 | 43.3 | 65.0 | **50.5** | 51.3 | +23.4 |

(b) **LLaVa OneVision.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *LLaVa OneVision 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 64.1 | 30.0 | 17.6 | 29.4 | 32.5 | 41.4 | 37.8 | 32.5 | 42.6 | 43.1 | 37.1 | |
| *Random Ctx* | 4 | 41.6 | 25.6 | 19.3 | 29.4 | 32.1 | 33.0 | 37.6 | 32.2 | 26.3 | 26.3 | 30.3 | -6.8 |
| | 8 | 40.4 | 25.2 | 20.4 | 29.4 | 32.0 | 33.3 | 37.7 | 32.2 | 26.1 | 26.4 | 30.3 | -6.8 |
| | 16 | 32.1 | 25.5 | 20.3 | 29.4 | 28.3 | 34.0 | 37.8 | 32.2 | 27.2 | 26.3 | 29.3 | -7.8 |
| *Pseudo ICL* | 4 | 23.3 | 22.5 | 16.3 | 29.4 | 27.1 | 41.9 | 35.1 | 30.2 | 37.8 | 20.3 | 28.4 | -8.7 |
| | 8 | 58.1 | 27.6 | 16.9 | 29.4 | 25.6 | 38.7 | 38.2 | 32.1 | 39.6 | 27.5 | 33.4 | -3.7 |
| | 16 | 57.3 | 27.0 | 17.9 | 29.4 | 25.3 | 39.8 | 38.2 | 31.6 | 41.4 | 30.6 | 33.8 | -3.3 |
| CIRCLE | 4 | 79.4 | 61.7 | 23.8 | 29.4 | **48.4** | 41.0 | 38.4 | **39.8** | 64.2 | 39.8 | 46.6 | +9.5 |
| | 8 | 81.1 | 52.3 | 32.9 | **29.8** | 40.6 | 43.9 | 39.3 | 34.8 | **66.5** | 53.8 | 47.5 | +10.4 |
| | 16 | **82.7** | **71.8** | **36.7** | **29.8** | 41.7 | **53.8** | **42.5** | 35.9 | 65.3 | **57.3** | **51.8** | +14.7 |

(c) **Phi-3.5-Vision and Phi-4-Multimodal.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phi-3.5-Vision* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 53.1 | 28.8 | 7.4 | 21.2 | 32.3 | 36.6 | 25.0 | 38.3 | 39.2 | 36.1 | 31.8 | |
| *Random Ctx* | 4 | 67.0 | 27.5 | 16.6 | 20.5 | 36.3 | 41.3 | 29.4 | 32.4 | 47.6 | 37.3 | 35.6 | +3.8 |
| | 8 | 62.0 | 25.3 | 24.3 | 22.9 | 36.6 | 37.6 | 29.8 | 31.7 | 44.5 | 32.8 | 34.8 | +3.0 |
| | 16 | 57.6 | 24.9 | 26.5 | 21.0 | 35.3 | 35.3 | 29.2 | 31.6 | 39.2 | 30.9 | 33.1 | +1.3 |
| *Pseudo ICL* | 4 | 47.4 | 25.3 | 6.7 | 20.8 | 27.3 | 32.7 | 29.2 | 44.3 | 32.8 | 32.4 | 29.9 | -1.9 |
| | 8 | 49.2 | 26.1 | 5.5 | 21.1 | 27.0 | 32.9 | 24.7 | 46.8 | 32.9 | 32.3 | 29.9 | -1.9 |
| | 16 | 52.1 | 25.5 | 5.8 | 20.4 | 27.0 | 32.9 | 28.7 | **46.9** | 34.0 | 32.9 | 30.6 | -0.2 |
| CIRCLE | 4 | 69.5 | 38.3 | **32.0** | 29.0 | **44.1** | 37.9 | 35.8 | 43.2 | 56.7 | **39.8** | **42.6** | +10.8 |
| | 8 | **72.7** | **39.4** | 30.0 | 28.4 | 42.5 | **49.4** | 35.9 | 36.3 | **57.2** | 29.3 | 42.1 | +10.3 |
| | 16 | 68.3 | 34.7 | 30.8 | **30.0** | 41.8 | 46.3 | **38.2** | 36.3 | 52.3 | 29.5 | 40.8 | +9.0 |
| *Phi-4-Multimodal* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 73.3 | 34.4 | 13.6 | 29.2 | 42.2 | 37.8 | 37.5 | 34.1 | 41.2 | 39.7 | 38.3 | |
| *Random Ctx* | 4 | 36.7 | 29.7 | 36.1 | 28.8 | 40.9 | 36.5 | 36.8 | 32.3 | 28.5 | 24.5 | 33.1 | -5.2 |
| | 8 | 39.5 | 27.8 | 36.9 | 29.1 | 41.4 | 36.4 | 37.6 | 32.3 | 28.9 | 24.1 | 33.4 | -4.9 |
| | 16 | 38.2 | 26.4 | 36.0 | 29.2 | 41.0 | 35.2 | 37.6 | 32.2 | 27.4 | 23.1 | 32.6 | -5.7 |
| *Pseudo ICL* | 4 | 75.5 | 32.8 | 14.9 | **29.3** | 42.7 | 43.8 | 38.2 | 34.5 | 47.4 | 32.5 | 39.2 | +0.9 |
| | 8 | 74.4 | 36.9 | 25.1 | **29.3** | 42.0 | 43.8 | 38.2 | 33.6 | 45.4 | 34.0 | 40.3 | +2.0 |
| | 16 | 76.4 | 28.9 | 13.2 | **29.3** | 41.9 | 43.0 | 38.2 | 34.7 | 46.7 | 31.5 | 38.4 | +0.1 |
| CIRCLE | 4 | **77.6** | 28.8 | 27.5 | 28.7 | 49.9 | 50.9 | 46.9 | 40.1 | 59.6 | **50.6** | 46.1 | +7.8 |
| | 8 | 71.7 | 38.7 | 36.6 | 28.4 | **53.8** | **61.7** | 54.7 | 47.0 | **59.8** | 28.7 | 48.1 | +9.8 |
| | 16 | 74.1 | **40.2** | **41.7** | 27.8 | 46.7 | 57.3 | **55.0** | **48.0** | 58.7 | 48.2 | **49.8** | +11.5 |

Table 9. **Open-world results.** ***Llama Inclusion*** (LI) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. Δ computed w.r.t. the average (Avg.) of *Zero-Shot*. For each LMM, **bold** indicates the best result. Results are split in Tabs. 9a to 9c for readability.

(a) **Qwen2-VL and Qwen2.5-VL.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 84.0 | 59.5 | 17.7 | 55.5 | 68.9 | 74.3 | 46.0 | 63.5 | 72.2 | 47.7 | 58.9 | |
| *Random Ctx* | 4 | 64.8 | 36.5 | 19.5 | 47.1 | 38.9 | 45.4 | 44.3 | 34.9 | 31.3 | 33.9 | 39.7 | -19.2 |
| | 8 | 63.2 | 24.9 | 20.7 | 45.9 | 31.7 | 39.7 | 41.5 | 28.0 | 33.1 | 26.2 | 35.5 | -23.4 |
| | 16 | 31.1 | 17.8 | 17.6 | 31.1 | 59.7 | 25.3 | 10.0 | 31.0 | 17.8 | 15.9 | 25.7 | -33.2 |
| *Pseudo ICL* | 4 | 85.8 | 60.5 | 9.9 | 53.6 | 52.3 | 70.9 | 36.2 | 44.6 | 76.4 | 58.1 | 54.8 | -4.1 |
| | 8 | 83.0 | 45.1 | 18.6 | 52.9 | 50.0 | 64.4 | 27.3 | 41.0 | 70.6 | 49.4 | 50.2 | -8.7 |
| | 16 | 82.6 | 47.6 | 6.4 | 55.5 | 52.1 | 62.5 | 29.7 | 44.1 | 64.8 | 51.3 | 49.7 | -9.2 |
| CIRCLE | 4 | 92.7 | **69.4** | **54.9** | 88.6 | **91.5** | 88.9 | **86.2** | 98.1 | 88.8 | 73.5 | **83.3** | +24.4 |
| | 8 | 93.5 | 68.3 | 32.2 | **92.4** | 91.0 | **93.8** | 75.0 | **98.8** | **90.6** | **85.9** | 82.2 | +23.3 |
| | 16 | **94.0** | 66.5 | 39.1 | 84.8 | 90.7 | 91.0 | 80.3 | 98.1 | 89.0 | 79.2 | 81.3 | +22.4 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 84.3 | 58.9 | 12.5 | 68.8 | 74.7 | 76.1 | 70.7 | 69.3 | 81.6 | 66.3 | 66.3 | |
| *Random Ctx* | 4 | 85.4 | 60.3 | 22.0 | 39.5 | 75.7 | 64.9 | 76.3 | 31.7 | 78.5 | 63.3 | 59.7 | -6.6 |
| | 8 | 86.1 | 61.4 | 21.7 | 37.6 | 73.7 | 65.4 | **77.1** | 31.2 | 78.1 | 62.5 | 59.5 | -6.8 |
| | 16 | 87.3 | 61.5 | 21.3 | 34.6 | 74.4 | 61.3 | 75.9 | 34.8 | 77.6 | 61.7 | 59.1 | -7.2 |
| *Pseudo ICL* | 4 | 84.3 | 55.3 | 10.7 | 38.1 | 65.9 | 61.0 | 64.4 | 41.3 | 76.9 | 61.9 | 56.0 | -10.3 |
| | 8 | 83.1 | 56.4 | 13.4 | 33.3 | 69.8 | 59.3 | 61.7 | 47.6 | 77.9 | 61.4 | 56.4 | -9.9 |
| | 16 | 83.0 | 56.3 | 11.3 | 30.5 | 69.3 | 61.1 | 66.1 | 41.3 | 78.4 | 61.2 | 55.9 | -10.4 |
| CIRCLE | 4 | 90.1 | 68.8 | 42.1 | 95.6 | 94.4 | 80.2 | 72.4 | 81.8 | 90.6 | 72.2 | 78.8 | +12.5 |
| | 8 | 89.3 | **77.4** | 40.1 | **96.3** | **95.4** | 78.6 | 67.4 | 78.8 | 90.0 | 58.0 | 77.1 | +10.8 |
| | 16 | **95.9** | 71.7 | **50.2** | 93.7 | 93.9 | **90.8** | 74.1 | **93.5** | **93.9** | **81.1** | **83.9** | +17.6 |

(b) **LLaVa OneVision.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *LLaVa OneVision 7B* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 81.3 | 45.6 | 11.8 | 68.9 | 48.9 | 22.0 | 50.2 | **84.4** | 25.0 | 27.0 | 46.5 | |
| *Random Ctx* | 4 | 24.3 | 6.6 | 9.9 | 67.9 | 18.0 | 4.5 | 40.0 | 80.4 | 3.6 | 5.3 | 26.0 | -20.5 |
| | 8 | 25.1 | 7.0 | 9.0 | 67.9 | 17.3 | 4.5 | 40.4 | 80.4 | 3.3 | 4.3 | 25.9 | -20.6 |
| | 16 | 21.8 | 16.3 | 7.8 | 0.0 | 6.5 | 2.6 | 0.0 | 0.0 | 0.7 | 0.3 | 5.6 | -40.9 |
| *Pseudo ICL* | 4 | 3.1 | 2.8 | 5.1 | 68.9 | 7.7 | 24.4 | 28.9 | 71.8 | 36.3 | 2.0 | 25.1 | -21.4 |
| | 8 | 87.7 | 61.3 | 15.1 | 67.9 | 43.9 | 11.8 | 46.1 | 80.8 | 23.2 | 13.4 | 45.1 | -1.4 |
| | 16 | 87.9 | 61.2 | 18.0 | 67.9 | 42.1 | 12.7 | 46.1 | 78.9 | 28.3 | 16.3 | 45.9 | -0.6 |
| CIRCLE | 4 | 85.5 | 84.8 | 34.3 | 96.0 | 93.2 | 46.9 | 70.4 | 38.8 | 84.2 | 38.4 | 67.2 | +20.7 |
| | 8 | 88.0 | 66.6 | 22.3 | **96.6** | 93.1 | 58.1 | 70.4 | 13.2 | 86.8 | 76.9 | 67.2 | +20.7 |
| | 16 | **96.5** | **88.9** | **66.8** | 65.4 | **95.1** | **73.0** | **72.6** | 64.4 | **89.8** | **80.9** | **79.3** | +32.8 |

(c) **Phi-3.5-Vision and Phi-4-Multimodal.**

| Method | Shots | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phi-3.5-Vision* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 75.0 | 45.6 | 1.7 | 60.7 | 61.8 | 42.9 | 47.4 | 47.7 | 46.4 | 38.8 | 46.8 | |
| *Random Ctx* | 4 | 66.6 | 19.9 | 5.0 | 39.2 | 46.5 | 20.0 | 20.7 | 64.2 | 31.9 | 19.2 | 33.3 | -13.5 |
| | 8 | 61.7 | 15.0 | 2.8 | 47.9 | 44.0 | 14.5 | 21.7 | 65.6 | 28.4 | 16.1 | 31.8 | -15.0 |
| | 16 | 56.5 | 15.4 | 0.8 | 55.6 | 43.4 | 8.6 | 20.2 | 60.9 | 24.1 | 15.0 | 30.0 | -16.8 |
| *Pseudo ICL* | 4 | 66.9 | 34.5 | 1.0 | 35.7 | 40.4 | 44.0 | 45.1 | 14.1 | 41.1 | 35.5 | 35.8 | -11.0 |
| | 8 | 68.4 | 39.8 | 0.0 | 35.8 | 31.7 | 40.0 | 31.8 | 15.8 | 50.3 | 32.5 | 34.6 | -12.2 |
| | 16 | 66.7 | 40.2 | 0.0 | 44.0 | 26.1 | 42.4 | 36.5 | 16.0 | 44.7 | 35.3 | 35.2 | -11.6 |
| CIRCLE | 4 | 79.2 | 66.0 | 29.5 | 98.5 | **89.9** | 78.1 | **86.4** | 79.2 | 82.4 | 61.4 | 75.1 | +28.3 |
| | 8 | 84.9 | **75.4** | 36.0 | 94.7 | 88.6 | 80.7 | 84.8 | 91.1 | 84.6 | **66.5** | 78.7 | +31.9 |
| | 16 | **93.5** | 75.1 | **47.8** | **99.1** | 84.1 | **94.6** | 85.7 | **100.0** | **90.6** | 52.1 | **82.3** | +35.5 |
| *Phi-4-Multimodal* | | | | | | | | | | | | | |
| *Zero-Shot* | - | 76.6 | 32.4 | 10.0 | 67.1 | 54.2 | 15.0 | 43.9 | 80.1 | 23.0 | 21.3 | 42.4 | |
| *Random Ctx* | 4 | 15.5 | 5.9 | 17.7 | 62.7 | 44.7 | 9.1 | 36.0 | 77.3 | 6.3 | 5.3 | 28.0 | -14.4 |
| | 8 | 19.7 | 3.3 | 15.4 | 64.1 | 45.3 | 8.1 | 40.3 | 79.6 | 6.2 | 4.5 | 28.6 | -13.8 |
| | 16 | 18.3 | 1.7 | 12.1 | 65.7 | 46.3 | 6.7 | 39.7 | 78.9 | 4.3 | 3.6 | 27.7 | -14.7 |
| *Pseudo ICL* | 4 | 75.7 | 26.8 | 5.4 | 67.7 | 50.6 | 16.6 | 45.9 | 77.8 | 28.1 | 13.0 | 40.8 | -1.6 |
| | 8 | 74.1 | 19.7 | 10.1 | 67.7 | 47.3 | 15.9 | 45.9 | 78.6 | 24.4 | 14.9 | 39.8 | -2.6 |
| | 16 | 76.6 | 42.7 | 2.4 | 67.8 | 45.5 | 15.6 | 45.9 | 75.5 | 26.6 | 12.2 | 41.1 | -1.3 |
| CIRCLE | 4 | 88.3 | 31.1 | 21.0 | **95.6** | 88.6 | **76.4** | **77.0** | 89.1 | 85.7 | 68.0 | 72.1 | +29.7 |
| | 8 | 87.6 | 72.1 | 38.7 | 88.5 | 89.0 | 75.5 | 73.3 | **94.4** | 83.1 | **73.7** | 77.6 | +35.2 |
| | 16 | **93.7** | **72.9** | **61.9** | 70.7 | **92.5** | 70.2 | 74.7 | 79.7 | **89.3** | 68.1 | **77.4** | +35.0 |

Table 10. **Streaming results.** We report results for *Llama Inclusion* (`LI`), *Semantic Similarity* (`SS`), *Concept Similarity* (`bCS`), and *Median Concept Similarity* (`mCS`). Purple indicates our CIRCLE. Higher is better on all metrics. For each LMM, **bold** indicates the best result.

| Method | Prototypical | | | | Non-prototypical | | | | Fine-grained | | | | Very fine-grained | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | `LI` | `SS` | `bCS` | `mCS` | `LI` | `SS` | `bCS` | `mCS` | `LI` | `SS` | `bCS` | `mCS` | `LI` | `SS` | `bCS` | `mCS` |
| *Qwen2-VL 7B* | | | | | | | | | | | | | | | | |
| *Zero-Shot* | 78.7 | 51.9 | **76.0** | 43.7 | 42.6 | 30.8 | **49.8** | 29.2 | 64.0 | 39.2 | 62.9 | 31.9 | 63.0 | 34.5 | 43.4 | 33.1 |
| *Pseudo ICL* | 65.9 | **62.9** | 70.5 | **61.4** | 36.5 | 36.5 | 44.3 | 36.3 | 54.0 | 54.0 | 59.3 | 52.6 | 40.0 | 40.0 | 46.5 | **39.9** |
| CIRCLE | **90.4** | 60.9 | 71.7 | 59.5 | **58.8** | **38.8** | 46.1 | **38.4** | **83.4** | **56.0** | **66.9** | **55.2** | **83.5** | **42.2** | **51.4** | 38.7 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | | | | | |
| *Zero-Shot* | 82.9 | 47.9 | **79.9** | 31.1 | 45.9 | 30.5 | **54.0** | 24.8 | 73.8 | 47.0 | **78.9** | 29.5 | 69.0 | 45.8 | **68.6** | 27.1 |
| *Pseudo ICL* | 63.7 | **70.4** | 70.6 | **70.4** | 21.7 | **40.2** | 40.3 | **40.1** | 50.2 | **58.2** | 58.3 | **58.0** | 73.6 | **45.9** | 46.2 | **44.1** |
| CIRCLE | **87.3** | 66.0 | 66.5 | 66.0 | **60.2** | 39.5 | 40.6 | 39.5 | **81.0** | 55.0 | 56.7 | 54.9 | **86.7** | 37.3 | 37.3 | 37.3 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | | | | | |
| *Zero-Shot* | 53.2 | 56.2 | 62.0 | 53.4 | 28.1 | 31.6 | 43.8 | 30.2 | **40.4** | 39.0 | 43.9 | 37.2 | **76.7** | 31.8 | 32.3 | 30.9 |
| *Pseudo ICL* | 44.8 | 45.0 | 56.0 | 39.0 | 25.2 | 25.8 | 41.1 | 23.1 | 25.7 | 34.1 | 42.3 | 33.3 | 72.4 | 30.8 | 30.8 | 30.8 |
| CIRCLE | **61.7** | **70.9** | **70.9** | **70.9** | **33.5** | **49.7** | **49.8** | **49.7** | 38.9 | **45.9** | **45.9** | **45.9** | 73.8 | **34.2** | **34.2** | **34.2** |
| *Phi-3.5-Vision* | | | | | | | | | | | | | | | | |
| *Zero-Shot* | 60.7 | 48.2 | **65.6** | 46.1 | 28.7 | 24.9 | **36.7** | 24.1 | 50.7 | 32.1 | **47.2** | 31.3 | 54.2 | 29.5 | **36.3** | 29.8 |
| *Pseudo ICL* | 49.5 | **58.8** | 61.0 | **58.5** | 19.3 | **29.1** | 35.5 | 28.9 | 41.3 | **42.7** | 45.6 | 42.0 | **68.8** | **32.5** | 32.5 | 32.5 |
| CIRCLE | **78.6** | 50.3 | 55.3 | 52.5 | **40.6** | 28.5 | 33.4 | **31.2** | **84.6** | 38.7 | 46.2 | **42.2** | 68.2 | 31.4 | 33.8 | **32.6** |
| *Phi-4-Multimodal* | | | | | | | | | | | | | | | | |
| *Zero-Shot* | 49.8 | 57.4 | 58.7 | 57.2 | 21.2 | 29.2 | 32.7 | 29.2 | 37.7 | 39.2 | 39.2 | 39.1 | 73.6 | 31.6 | 31.7 | 31.6 |
| *Pseudo ICL* | 50.6 | 62.3 | 62.3 | 62.3 | 18.8 | 36.1 | 36.7 | 36.0 | 36.7 | 41.8 | 41.8 | 41.8 | 71.3 | 31.7 | 31.7 | 31.7 |
| CIRCLE | **84.9** | **62.7** | **69.4** | **64.6** | **64.8** | **40.7** | **44.5** | **40.7** | **76.7** | **53.2** | **57.4** | **53.5** | **77.8** | **40.0** | **45.0** | **39.1** |

Table 11. **Streaming results.** ***Semantic Similarity*** (SS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. Δ computed w.r.t. the average (Avg.) of *Zero-Shot*. For each LMM, **bold** indicates the best result.

| Method | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 55.8 | 28.6 | 20.7 | 20.6 | 41.7 | 50.7 | 25.1 | 48.3 | 48.1 | 43.1 | 38.3 | |
| *Pseudo ICL* | **70.0** | 31.9 | 24.3 | 25.8 | 51.6 | **70.0** | 40.5 | **54.2** | 55.7 | **53.4** | 47.7 | +9.4 |
| CIRCLE | 63.2 | **34.0** | **32.1** | **33.8** | **54.1** | 62.0 | **51.9** | 50.6 | **58.5** | 50.2 | **49.0** | +10.7 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 48.8 | 28.3 | 19.0 | 36.7 | 47.4 | 52.4 | 41.1 | 54.9 | 47.0 | 44.2 | 42.0 | |
| *Pseudo ICL* | **78.7** | **37.1** | 27.4 | 29.8 | **66.5** | **66.8** | 41.2 | **61.9** | **62.2** | **56.2** | **52.8** | +10.8 |
| CIRCLE | 70.5 | 35.5 | **32.0** | **30.1** | 63.6 | 51.9 | **49.5** | 44.4 | 61.6 | 51.0 | 49.0 | +7.0 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 68.9 | 32.0 | 19.4 | **29.4** | 37.5 | 41.6 | 37.8 | 34.3 | 43.4 | 43.4 | 38.8 | |
| *Pseudo ICL* | 59.7 | 29.0 | 17.0 | **29.4** | 27.6 | 36.8 | 37.9 | 32.2 | 30.4 | 31.5 | 33.1 | -5.7 |
| CIRCLE | **77.4** | **57.8** | **39.7** | **29.4** | **42.5** | **54.2** | **40.9** | **39.1** | **64.3** | **51.7** | **49.7** | +10.9 |
| *Phi-3.5-Vision* | | | | | | | | | | | | |
| *Zero-Shot* | 53.2 | 29.1 | 7.4 | 19.9 | 31.6 | 40.2 | 24.6 | **39.1** | 43.2 | 38.3 | 32.6 | |
| *Pseudo ICL* | **71.7** | **32.8** | 11.8 | **29.3** | 44.1 | **50.7** | **33.2** | 35.7 | **45.8** | **42.7** | **39.8** | +7.2 |
| CIRCLE | 55.4 | 28.1 | **24.4** | 26.6 | **44.2** | 39.7 | 32.1 | 36.3 | 45.1 | 33.0 | 36.5 | +3.9 |
| *Phi-4-Multimodal* | | | | | | | | | | | | |
| *Zero-Shot* | 73.5 | 33.9 | 13.8 | 29.2 | 42.2 | 37.8 | 37.5 | 34.1 | 41.2 | 40.0 | 38.3 | |
| *Pseudo ICL* | **78.1** | 40.5 | 28.6 | **29.4** | 42.4 | 44.9 | 38.1 | 34.0 | 46.5 | 39.3 | 42.2 | +3.9 |
| CIRCLE | 69.1 | **41.4** | **34.3** | 27.7 | **49.0** | **60.5** | **50.2** | **52.3** | **56.3** | **46.5** | **48.7** | +10.4 |

Table 12. **Streaming results.** ***Concept Similarity*** (bCS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. $\Delta$ computed w.r.t. the average (Avg.) of *Zero-Shot*. For each LMM, **bold** indicates the best result.

| Method | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | **81.3** | **50.3** | **39.8** | 30.7 | **68.7** | 77.0 | 43.2 | 55.7 | **70.7** | 59.0 | 57.6 | |
| *Pseudo ICL* | 79.9 | 44.2 | 32.8 | 30.7 | 59.1 | **77.3** | 41.4 | **62.3** | 61.0 | 55.8 | 54.5 | -3.1 |
| CIRCLE | 77.0 | 42.7 | 36.2 | **46.9** | 62.6 | 71.7 | **66.5** | 56.0 | 66.3 | **59.3** | **58.5** | +0.9 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | **85.6** | **53.4** | **41.3** | **68.7** | **79.7** | **79.6** | **77.3** | **68.5** | **74.2** | **67.2** | **69.5** | |
| *Pseudo ICL* | 79.1 | 37.2 | 27.5 | 30.3 | 66.9 | 66.9 | 41.2 | 62.2 | 62.2 | 56.2 | 53.0 | -16.5 |
| CIRCLE | 71.4 | 37.5 | 32.1 | 30.2 | 64.7 | 52.5 | 52.9 | 44.4 | 61.7 | 52.2 | 50.0 | -19.5 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 79.1 | 46.9 | 41.0 | **29.4** | **51.9** | 41.9 | 37.9 | 35.3 | 44.8 | 43.5 | 45.2 | |
| *Pseudo ICL* | **81.6** | 49.4 | 36.9 | **29.4** | 48.6 | 40.4 | 37.9 | 32.2 | 31.5 | 37.0 | 42.5 | -2.7 |
| CIRCLE | 77.4 | **57.8** | **39.7** | **29.4** | 42.5 | 54.2 | **40.9** | **39.1** | **64.3** | **51.8** | **49.7** | +4.5 |
| *Phi-3.5-Vision* | | | | | | | | | | | | |
| *Zero-Shot* | 73.6 | **43.5** | 16.4 | 29.6 | 44.9 | **58.1** | **38.7** | **43.0** | **57.6** | **50.1** | **45.5** | |
| *Pseudo ICL* | **74.1** | 39.9 | 20.2 | 29.4 | 44.7 | 56.2 | 35.9 | 35.7 | 47.9 | 46.3 | 43.0 | -2.5 |
| CIRCLE | 59.3 | 34.8 | **24.9** | **30.9** | **53.6** | 46.7 | 38.3 | 36.6 | 51.3 | 40.5 | 41.7 | -3.8 |
| *Phi-4-Multimodal* | | | | | | | | | | | | |
| *Zero-Shot* | 75.8 | 40.6 | 17.4 | 29.3 | 42.2 | 37.9 | 37.5 | 34.1 | 41.7 | 40.1 | 39.6 | |
| *Pseudo ICL* | **78.2** | 40.6 | 30.0 | 29.4 | 42.4 | 44.9 | 38.1 | 34.0 | 46.5 | 39.4 | 42.4 | +2.8 |
| CIRCLE | 77.9 | **44.0** | **35.3** | **34.5** | **53.7** | **64.3** | **54.2** | **55.5** | **61.0** | **54.1** | **53.4** | +13.8 |

Table 13. **Streaming results.** ***Median Concept Similarity*** (mCS) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. $\Delta$ computed w.r.t. the average (Avg.) of *Zero-Shot*. For each LMM, **bold** indicates the best result.

| Method | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 51.3 | 27.9 | 20.6 | 20.4 | 33.3 | 37.3 | 25.0 | 45.8 | 36.1 | 38.9 | 33.7 | |
| *Pseudo ICL* | **69.4** | 31.2 | 24.9 | 25.7 | 50.1 | **67.2** | 40.5 | **54.2** | 53.3 | **52.9** | 46.9 | +13.2 |
| CIRCLE | 61.4 | **34.8** | **31.1** | **30.0** | **54.9** | 61.4 | **49.3** | 47.5 | **57.5** | 49.2 | **47.7** | +14.0 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 33.0 | 26.3 | 18.6 | 24.7 | 29.5 | 31.8 | 27.3 | 29.4 | 29.3 | 29.6 | 27.9 | |
| *Pseudo ICL* | **78.7** | **37.0** | 27.2 | 29.8 | **66.2** | **66.5** | 41.2 | **58.4** | **62.1** | **55.9** | **52.3** | +24.4 |
| CIRCLE | 70.5 | 35.8 | **31.8** | **30.2** | 63.1 | 52.1 | **49.6** | 44.4 | 61.6 | 51.0 | 49.0 | +21.1 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 64.1 | 30.0 | 17.6 | **29.4** | 32.5 | 41.4 | 37.8 | 32.5 | 42.6 | 43.1 | 37.1 | |
| *Pseudo ICL* | 48.3 | 26.8 | 16.5 | **29.4** | 26.5 | 35.5 | 37.9 | 32.2 | 29.7 | 26.1 | 30.9 | -6.2 |
| CIRCLE | **77.4** | **57.8** | **39.7** | **29.4** | **42.5** | **54.2** | **40.9** | **39.0** | **64.3** | **51.5** | **49.7** | +12.6 |
| *Phi-3.5-Vision* | | | | | | | | | | | | |
| *Zero-Shot* | 53.1 | 28.8 | 7.4 | 21.2 | 32.3 | 36.6 | 25.0 | **38.3** | 39.2 | 36.1 | 31.8 | |
| *Pseudo ICL* | **71.7** | **32.6** | 11.8 | **29.4** | 44.0 | **48.9** | 33.1 | 35.6 | 45.3 | **42.3** | **39.5** | +7.7 |
| CIRCLE | 56.7 | 31.5 | **24.6** | 28.8 | **44.2** | 44.8 | **37.8** | 36.4 | **48.2** | 37.6 | 39.1 | +7.3 |
| *Phi-4-Multimodal* | | | | | | | | | | | | |
| *Zero-Shot* | 73.3 | 34.4 | 13.6 | 29.2 | 42.2 | 37.8 | 37.5 | 34.1 | 41.2 | 39.7 | 38.3 | |
| *Pseudo ICL* | **78.1** | 40.6 | 28.6 | **29.4** | 42.3 | 44.9 | 38.1 | 34.0 | 46.5 | 39.0 | 42.1 | +3.8 |
| CIRCLE | 71.1 | **42.2** | **33.6** | 27.4 | **48.9** | **60.2** | **51.5** | **50.7** | **58.1** | **46.3** | **49.0** | +10.7 |

Table 14. **Streaming open-world results.** ***Llama Inclusion*** (LI) on the ten datasets. Purple indicates our CIRCLE. Higher is better. For each LMM, **bold** indicates the best result. $\Delta$ computed w.r.t. the average (Avg.) of *Zero-Shot*. For each LMM, **bold** indicates the best result.

| Method | C101 | DTD | ESAT | FGVC | FLWR | FOOD | PETS | CARS | S397 | U101 | Avg. | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen2-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 84.0 | 59.5 | 17.7 | 55.5 | 68.9 | 74.3 | 46.0 | 63.5 | 72.2 | 47.7 | 58.9 | |
| *Pseudo ICL* | 83.9 | 41.8 | 12.6 | 63.6 | 55.8 | 60.3 | 49.0 | 42.8 | 47.8 | 33.7 | 49.1 | -9.8 |
| CIRCLE | **93.5** | **61.1** | **43.9** | **76.4** | **90.0** | **85.5** | **74.7** | **90.7** | **87.2** | **71.5** | **77.4** | +18.5 |
| *Qwen2.5-VL 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 84.3 | 58.9 | 12.5 | 68.8 | 74.7 | **76.1** | 70.7 | 69.3 | 81.6 | 66.3 | 66.3 | |
| *Pseudo ICL* | 80.9 | 19.1 | 12.7 | 72.8 | 62.8 | 40.2 | 47.8 | 74.4 | 46.5 | 33.5 | 49.0 | -17.3 |
| CIRCLE | **89.5** | **60.3** | **42.9** | **94.2** | **93.5** | 73.5 | **76.1** | **79.1** | **85.0** | **77.3** | **77.2** | +10.9 |
| *LLaVa OneVision 7B* | | | | | | | | | | | | |
| *Zero-Shot* | 81.3 | 45.6 | 11.8 | **68.9** | **48.9** | 22.0 | **50.2** | **84.4** | 25.0 | **27.0** | 46.5 | |
| *Pseudo ICL* | **81.5** | **47.5** | 3.9 | 65.9 | 25.7 | 10.1 | 41.4 | 79.0 | 8.1 | 24.3 | 38.7 | -7.8 |
| CIRCLE | 77.0 | 40.3 | **37.7** | 65.7 | 45.4 | **23.5** | 47.8 | 81.9 | **46.4** | 22.4 | **48.8** | +2.3 |
| *Phi-3.5-Vision* | | | | | | | | | | | | |
| *Zero-Shot* | 75.0 | 45.6 | 1.7 | 60.7 | 61.8 | 42.9 | 47.4 | 47.7 | 46.4 | 38.8 | 46.8 | |
| *Pseudo ICL* | 69.4 | 31.7 | 4.8 | 60.9 | 56.9 | 31.3 | 35.8 | **76.7** | 29.5 | 21.5 | 41.9 | -4.9 |
| CIRCLE | **79.9** | **46.6** | **12.0** | **96.0** | **88.0** | **87.0** | **78.9** | 40.3 | **77.2** | **63.1** | **66.9** | +20.1 |
| *Phi-4-Multimodal* | | | | | | | | | | | | |
| *Zero-Shot* | 76.6 | 32.4 | 10.0 | 67.1 | 54.2 | 15.0 | 43.9 | **80.1** | 23.0 | 21.3 | 42.4 | |
| *Pseudo ICL* | 76.1 | 22.7 | 17.4 | 65.9 | 47.6 | 16.9 | 45.7 | 76.7 | 25.1 | 16.3 | 41.0 | -1.4 |
| CIRCLE | **86.4** | **65.1** | **60.5** | **84.1** | **86.4** | **78.6** | **65.0** | 71.6 | **83.5** | **68.6** | **75.0** | +32.6 |

**Ground Truth: airplane**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The object in the photo is an airplane. | helicopter | The object in the image is an airplane. | airplane, jet, airliner |

**Ground Truth: camera**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The object in the photo is a Canon camera. | laptop | The object in the image is a Canon camera. | camera, point-and-shoot, Canon |

**Ground Truth: ferry**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The object in the photo is a boat. | airplane 29 1 17 | The object in the image is a boat. | boat, ferry, vessel |

Figure 7. **Qualitative results from Caltech101 [14].** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE . We use Qwen2-VL [37] 7B as the LMM.

**Ground Truth: perforated**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The photo shows a black background with a grid of small, evenly spaced dots. This pattern is often referred to as a "dot grid" or "dotted grid." | The object in the photo is a piece of fabric with a striped pattern. | The image shows a piece of wood with a visible knot and a crack. The wood appears to be a wooden board or panel, and the knot and crack are natural features of the wood. | grid, perforated, metal |

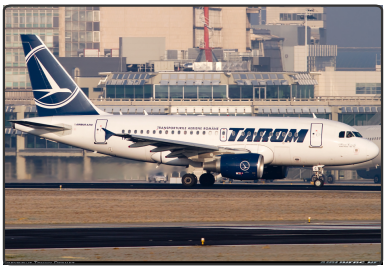
**Ground Truth: pleated**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The object in the photo is a piece of fabric with a pleated texture. The pleats create a wave-like pattern, giving the fabric a distinctive and textured appearance. | The object in the photo is a woven basket. | The object in this image is a cluster of dry, brown grasses. The grasses appear to be in a state of dormancy, possibly due to seasonal changes or lack of water. The image shows the grasses in a natural outdoor setting, likely a field or meadow. | textile, pleated, fabric |

**Ground Truth: zigzagged**

| VANILLA | RANDOM | PSEUDO-LABEL | CIRCLE |
|---|---|---|---|
| The object in the photo is a piece of fabric or clothing with a zigzag pattern. The pattern consists of alternating light and dark zigzag lines, creating a gradient effect from top to bottom. | The object in the photo is a waffle. | The image shows a pattern of zigzag lines in alternating colors of black and white. This type of pattern is often used in textiles, wallpaper, or other decorative surfaces. The zigzag design can be found in various forms of art and design, and it is known for its geometric and visually striking appearance. | pattern, chevron, zigzag |

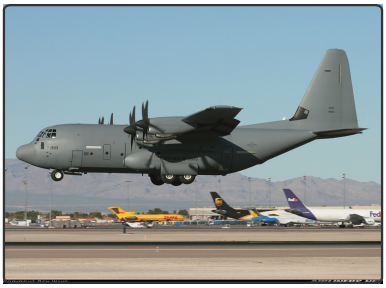
Figure 8. **Qualitative results from DTD [9]** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE . We use Qwen2-VL [37] 7B as the LMM.




**Ground Truth: A318**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is an airplane. | airplane |

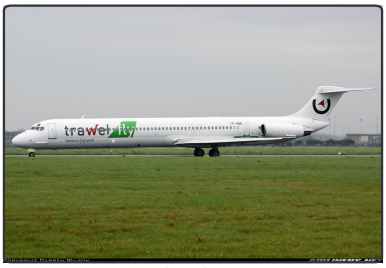
| PSEUDO-LABEL | CIRCLE |
|---|---|
| The object in the image is an airplane. | airliner, commercial, Airbus A318, Tarom, Romanian airline |



**Ground Truth: C-130**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is an aircraft. | The object in the photo is an airplane. |

| PSEUDO-LABEL | CIRCLE |
|---|---|
| The object in the image is an airplane. | transport aircraft, military, C-130 Hercules |

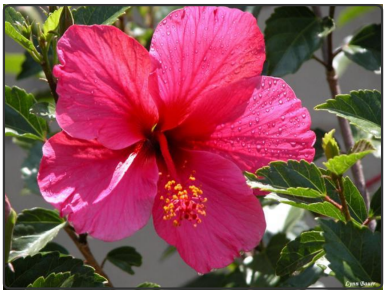



**Ground Truth: MD-80**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is an airplane. | The object in the photo is an airplane. |

| PSEUDO-LABEL | CIRCLE |
|---|---|
| The object in the image is an airplane. | airliner, commercial, MD-80 |

Figure 9. **Qualitative results from FGVC Aircraft [24].** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE. We use Qwen2-VL [37] 7B as the LMM.



**Ground Truth: hibiscus**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a hibiscus flower. | flower |

| PSEUDO-LABEL | CIRCLE |
|---|---|
| The image shows a cluster of pink flowers with white centers. The flowers appear to be in full bloom and are likely part of a garden or natural setting. | hibiscus, hibiscus, pink hibiscus |

**Ground Truth: globe thistle**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a flower. It appears to be a type of thistle, characterized by its spiky, round, and blue-purple petals. | The object in the photo is a flower. |

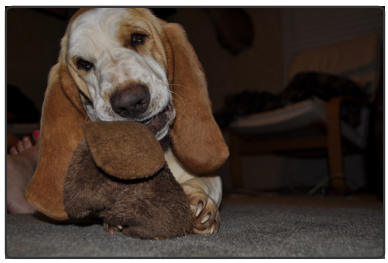
| PSEUDO-LABEL | CIRCLE |
|---|---|
| The image shows a group of passionflowers. Passionflowers are a type of flowering plant in the genus Passiflora, known for their large, showy flowers that often have a distinctive shape resembling a cross. The flowers in the image have vibrant purple petals with a white center and long, white stamens. | blue flower, echinops, globe thistle |

**Ground Truth: bird of paradise**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a bird of paradise flower (Strelitzia reginae). It is characterized by its vibrant, colorful flowers that resemble the wings of a bird. The flower has bright yellow and blue petals with a red base. | The object in the photo is a flower. |

| PSEUDO-LABEL | CIRCLE |
|---|---|
| The image shows a pink ginger flower (Alpinia officinarum). Ginger flowers are known for their large, showy blooms that often have a pink or red coloration. They are typically found in tropical and subtropical regions and are often used in ornamental gardens. The flower in the image has a distinctive shape | bird of paradise, tropical flower |

Figure 10. **Qualitative results from Flowers102 [27].** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE. We use Qwen2-VL [37] 7B as the LMM.

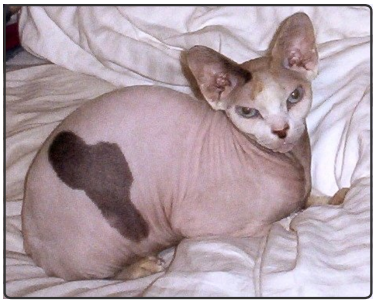


**Ground Truth: basset hound**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a toy. The dog is holding it in its mouth. | The object in the photo is a toy. |
| **PSEUDO-LABEL** | **CIRCLE** |
| The image shows a black cat lying down on a white surface. | dog, basset hound, toy |

**Ground Truth: english setter**

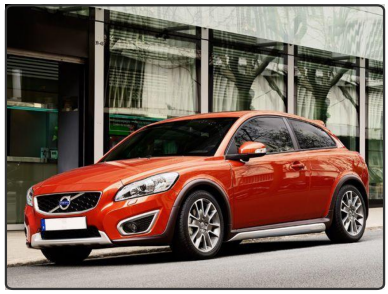
| VANILLA | RANDOM |
|---|---|
| The object in the photo is a dog. | The object in the photo is a dog. |
| **PSEUDO-LABEL** | **CIRCLE** |
| The object in the image is a cat. | dog, English Setter |



**Ground Truth: sphynx**

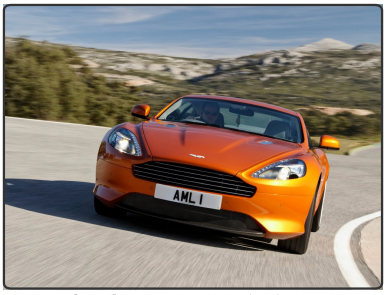
| VANILLA | RANDOM |
|---|---|
| The object in the photo is a cat. | cat |
| **PSEUDO-LABEL** | **CIRCLE** |
| The image shows a white dog with a fluffy coat, lying on a wooden surface with fallen leaves around it. The dog appears to be a Samoyed or a similar breed. | cat, hairless, Sphynx |

Figure 11. **Qualitative results from Oxford Pets [28].** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE. We use Qwen2-VL [37] 7B as the LMM.



**Ground Truth: 2012 Volvo C30 Hatchback**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a car. Specifically, it appears to be a Volvo model, as indicated by the logo on the front grille. | The object in the photo is a car. |
| **PSEUDO-LABEL** | **CIRCLE** |
| The object in the image is a gray pickup truck. | coupé, Volvo, Volvo C30 |



**Ground Truth: 2012 Aston Martin Virage Coupe**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a sports car. | The object in the photo is a car. |
| **PSEUDO-LABEL** | **CIRCLE** |
| The object in the image is a black Hyundai Tucson SUV. | sports car, luxury sports car, Aston Martin sports car |

**Ground Truth: 2012 Chevrolet Silverado 1500 Hybrid Crew Cab**

| VANILLA | RANDOM |
|---|---|
| The object in the photo is a silver Chevrolet Silverado pickup truck. | The object in the photo is a 2012 Chevrolet Silverado 1500 Regular Cab pickup truck. |
| **PSEUDO-LABEL** | **CIRCLE** |
| The object in the image is a black sedan parked in a lot. | truck, pickup truck, Chevrolet Silverado |

Figure 12. **Qualitative results from Stanford Cars [19].** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE. We use Qwen2-VL [37] 7B as the LMM.

**Ground Truth: lighthouse**

**VANILLA**

The object in the photo is a lighthouse.

**RANDOM**

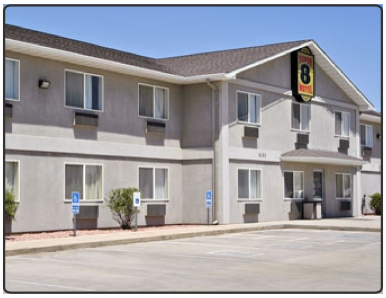
The object in the photo is a lighthouse.

**PSEUDO-LABEL**

The image shows a building with a large "T" logo on the front, two garage doors, and a small awning on the right side. The building appears to be a garage or a storage facility.

**CIRCLE**

lighthouse, lighthouse, coast



**Ground Truth: motel**

**VANILLA**

The object in the photo is a hotel or motel building.

**RANDOM**

The object in the photo is a building, specifically a hotel.

**PSEUDO-LABEL**

The object in the image is a pizza-making machine. It is a large, industrial machine used in a pizza factory or production facility. The machine is designed to automate the process of making pizzas, including the stretching of the dough and the application of toppings. The image shows a person working with the machine, likely adding ingredients

**CIRCLE**

hotel, motel

**Ground Truth: putting green**

**VANILLA**

The photo shows a small, rectangular putting green with a red flag in the center. The putting green is surrounded by a border of white rocks and a bed of small rocks and flowers.

**RANDOM**

golf course

**PSEUDO-LABEL**

The image shows a cityscape with a prominent tall tower, which is Coit Tower, located in San Francisco, California. The tower is situated on top of Telegraph Hill and is surrounded by various buildings and trees. The sky is clear, and the overall scene captures the urban landscape of the city.

**CIRCLE**

golf course, putting green

Figure 13. **Qualitative results from SUN397 [38]** We visualize three distinct samples, showing the response given by the *Vanilla* model, with *Random* context, with *Pseudo-labeling* examples, and using our CIRCLE. We use Qwen2-VL [37] 7B as the LMM.

## References


[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*, 2024. 1, 4, 7, 5

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv*, 2023. 1, 7

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4, 7, 5

[4] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *CVPR-WS*, 2024. 4

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 4, 1

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv*, 2020. 2

[7] 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 *CVPR*, 2024. 1

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 2

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 4, 1, 18

[10] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *NeurIPS*, 2023. 2, 4, 5, 6, 7, 1

[11] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification and semantic segmentation. *arXiv*, 2024. 2

[12] Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. *ICCV*, 2025. 1, 2, 4, 6, 5

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 4, 5

[14] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 4, 1, 18

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv*, 2023. 1, 2

[16] Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, and Massimiliano Mancini. Compositional caching for training-free open-vocabulary attribute detection. In *CVPR*, 2025. 2, 3

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 4, 1

[18] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024. 8

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV-WS*, 2013. 4, 1, 20

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv*, 2024. 1, 4, 7, 5

[21] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. Revisiting mllms: An in-depth analysis of image classification abilities. *arXiv*, 2024. 1, 2

[22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 1, 2

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv*, 2024. 2

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, 2013. 4, 1, 19

[25] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin

Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. 4, 7, 5

[26] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv*, 2022. 5

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*, 2008. 4, 1, 19

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 4, 1, 20

[29] Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multi-modal in-context learning? an in-depth exploration. *NeurIPS*, 2024. 2, 3

[30] 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 *ICML*, 2021. 1, 2, 4, 5

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019. 6

[32] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 2022. 4

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 4, 1

[34] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 2

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2

[36] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, 2023. 2, 3

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024. 1, 4, 8, 5, 18, 19, 20, 21

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 4, 1, 21

[39] Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. Object recognition as next token prediction. In *CVPR*, 2024. 1, 2

[40] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 1, 2

[41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 2

[42] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *NAACL*, 2025. 1, 2

[43] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv*, 2021. 2, 3, 4, 5, 1

[44] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv*, 2024. 2, 3

[45] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *NeurIPS*, 2023. 2, 3

[46] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *NeurIPS*, 2024. 1, 2

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 4