A. Supplementary Material

In the following, we provide additional information on our analyses. First, we report further detail on the considered datasets, models (A.1), and metrics (A.2), followed by the extended results of each model for each dataset (A.3). Then, we extend our main analyses by evaluating with an Elo ranking system which model provides the best responses, using a Llama instance to score wins. We continue the analysis by evaluating the percentage of agreement between models for the three prediction groups not present in the main paper, and we use RAM++ to tag images by checking whether we can improve the model performance by using prompts that foster multi-label responses, *e.g.*, listing the objects in the scene, or describing the image (A.4). Finally, we report additional tables and visualizations to accompany the studies in the main manuscript (A.5).

A.1. Additional details on the datasets and models

The datasets used in our evaluation are summarized in Tab. 3. For the experiments we used the same training and test splits used in previous works [17], while a summary of the LMMs used in this study and their differences is in Tab. 4.

A.2. Additional details on the metrics

For computing the inclusion metric, we instruct Llama 3.2 [25] to score good and bad LMM responses with the following prompt:

Llama inclusion instruction

You are a model that determines whether an answer is a good reply to a question given also its target value.

This is the question: What type of object is in this image?
This is the answer: %s
This is the target value: %s

If the answer describes the target, reply positively. If the answer includes the target value or a synonym of it, reply positively. If the target is generic but it is related to the answer, reply positively. Reply only with "1" if yes, or "0" if no.

A.3. Extended results

We report the per-dataset results of the evaluated LMMs, split into one table for each of the considered metrics, *i.e.*, text inclusion in Tab. 5, Llama inclusion in Tab. 6, semantic similarity in Tab. 7, and concept similarity in Tab. 8.

Dataset	Images	Classes
CALTECH101 [23] (C101)	2,465	100
DTD [16]	1,692	47
EUROSAT [26] (ESAT)	8,100	10
FGVCAIRCRAFT [44] (FGVC)	3,333	100
FLOWERS102 [46] (FLWR)	2,463	102
FOOD101 [8] (FOOD)	30,300	101
OXFORDPETS [47] (PETS)	3,669	37
STANFORD CARS [33] (CARS)	8,041	196
SUN397 [62] (S397)	19,850	397
UCF101 [56] (U101)	3,783	101

Table 3. Summary details of the datasets used in our analyses.

A.4. Additional analyses

Which model provides the best responses? To analyze which model provides the best responses, we compare their generations in pairs. Specifically, for each of the ten datasets, we randomly sample 10'000 pairs of generations, and instruct a Llama 3.2 model to identify the best response in the pair, similarly to what done in the Chatbot Arena [15] but through automatic evaluation with LLM-as-a-judge [71]. We use the following prompt to instruct Llama 3.2 to judge the pairs of predictions and decide for a win:

Llama Elo ranking

You are a model that discriminates whether labels A or B better align with a target value.

This is label A: %s
This is label B: %s
This is the target value: %s

Does A align better with the target value? Does B align better with the target value? Reply only with "1" if A wins over B, or "0" if B wins over A.

We directly compare the quality of the outputs by evaluating the Elo score [21] of these model responses and report the average on the ten datasets in Tab. 9. Results show that Qwen2VL models are the best at providing accurate predictions, similar to the trend in Tab. 1.

Which models agree the most with each other? To complement the analysis of the main paper, here we show the pair-wise agreement on the model predictions on group beyond the correct and specific one, showing the results in Fig. 9 for correct but generic, Fig. 10 for wrong but specific, and Fig. 11 for wrong and generic. The trends follow those of the main paper (Fig. 5) *i.e.*, where models of the same families tend to agree on the same samples, generalizing those findings across groups.

Model	Vision Enc	Language Enc	Training	Pre-training
IDEFICS2 [34]	SOViT (SigLIP), 0.4B params; max 980x980.	Mistral 7B	Interleaved web docs, image-caption pairs (LAION-COCO), OCR data; fine-tuned on 50 curated datasets.	Joint dual encoder training with Perceiver pooling for vision-text alignment.
INSTRUCTBLIP [19]	ViT-g (BLIP-2), 1.1B params; 224x224.	Vicuna 7B	26 datasets transformed into instruction-tuning format: captioning, VQA, image generation.	Two-stage pre-training: Vision-language alignment via BLIP-2 and instruction-aware Query Transformer for task-specific feature extraction.
INTERNVL2 [13]	InternViT (custom), 0.3B params (or 6B for larger models); dynamic resolution, max 40 tiles of 448×448.	Qwen2 0.5B (for 1B and 2B versions), or InternLM2 8B (for 8B version).	Interleaved image-text, multilingual OCR, mathematical charts; strict quality control.	Progressive training: masked video modeling, cross-modal contrastive learning, and next-token prediction with spatiotemporal focus.
LLAVA-1.5 [41]	ViT-L (CLIP), 0.3B params; 336x336.	Vicuna 7B	158K multimodal instruction-following samples; pre-trained on filtered CC dataset (596K image-text pairs).	Frozen vision encoder during feature alignment stage; end-to-end fine-tuning.
LLAVA-NEXT [36]	ViT-L (CLIP), 0.3B params; 336x336, 672x672, 336x1344, and 1344x336.	Mistral 7B, or Vicuna 7B	Diverse tasks, including multi-image and video understanding.	Builds on LLaVA with extended ViT and additional multimodal datasets for improved generalization.
LLAVA-OV [37]	SOViT (SigLIP), 0.4B params; dynamic resolution (AnyRes-9), max 2304x2304.	Qwen2 0.5B, or Qwen2 7B	Single-image and video scenarios with task transfer capabilities; diverse visual benchmarks.	Pre-trained with balanced visual token representation across scenarios to enable task transfer.
PHI-3-VISION [1]	ViT-L (CLIP), 0.4B params; dynamic resolution, max 1344x1344.	Phi-3 Mini (3.8B params)	Synthetic data, filtered public docs, high-quality interleaved text-image data, math/code examples.	Multi-stage training: custom vision encoder aligned with Phi-3 Mini language model using interleaved and fine-grained tasks.
QWEN2VL [60]	ViT (custom), 0.6B params; dynamic resolution (Naive Dynamic Resolution), no max.	Qwen2 1.5B, or Qwen2 7B	Multilingual datasets: MathVista, DocVQA, RealWorldQA; supports videos (20+ min) and multilingual text in images.	Pre-trained with dynamic resolution ViT for flexible input sizes and multilingual alignment strategies.

Table 4. Summary details of the Language Multimodal Models used in our analyses.

Predicting more concepts. The experiment using RAM++ to tag images suggested that LMMs often fail to predict the class names because they focus on the wrong part of the image. However, when prompted to provide multiple candidates, do LMMs get the correct prediction? To investigate this, we ask the model to (i) list the objects in the image; (ii) caption it, or (iii) describe its content. We report the relative gain per model in Tab. 12. The results show that providing outputs that focus on multiple labels on average improves the concept-based similarity, with the only exception of the caption case. Text inclusion improves consistently, showing that predictions become correct even according to this strict metric. Overall, these results highlight how LMM mistakes can be ascribed by mismatches between the label and the focus of the annotator, with the models often focusing on grounded image content even in case of mistakes.

Larger models. In Tab. 10, we add 6 larger models (green) to the original 13 base, with scales from 13B to 72B. Notably, *scaling has mixed impacts*, sometimes leading to better performance (*e.g.*, InternVL2 26B, Qwen2-VL 72B) and sometimes worse (*e.g.*, InstructBLIP 13B, LLaVA-NeXT 34B). Particularly, the language encoder in LLaVA-NeXT changes between 13B (Vicuna) and 34B (Yi), highlighting that *the pre-training data has a stronger effect than scaling*.

Commercial models. In Tab. 11, we report results for commercial models on a subset of the considered datasets. We compare these models against all the previously considered

Model	Textual inclusion										
	C101	DTD	ESAT	FGVC	FLWR	FOOD	PETS	CARS	S397	U101	Avg.
IDEFICS2 [34] 8B	52.0	1.7	1.6	0.0	0.8	8.2	0.1	0.0	9.6	7.9	8.2
INSTRUCTBLIP [19] Vicuna 7B	47.8	3.0	5.5	0.0	6.0	24.3	0.8	0.0	11.6	9.6	10.9
INTERNVL2 [12, 13] 2B	52.8	10.8	7.4	1.4	14.1	23.3	7.2	0.0	21.1	12.4	15.0
INTERNVL2 [12, 13] 4B	49.6	11.8	6.0	3.4	12.8	28.2	7.8	0.0	23.0	12.7	15.5
INTERNVL2 [12, 13] 8B	55.0	12.5	6.0	4.6	19.1	33.9	13.8	**0.1**	26.3	**14.4**	18.6
LLAVA-1.5 [41] 7B	51.6	6.0	**11.7**	0.1	6.7	17.6	1.1	0.0	17.6	8.2	12.1
LLAVA-NEXT [36] (Mistral 7B)	58.0	13.6	7.4	2.8	17.6	35.5	27.1	0.0	25.4	13.0	20.0
LLAVA-NEXT [36] (Vicuna 7B)	54.9	12.2	7.2	2.5	11.9	29.6	9.4	0.0	24.0	12.5	16.4
LLAVA-OV [37] (Qwen2 0.5B)	53.4	9.2	4.2	1.2	2.9	12.6	2.5	**0.1**	15.5	8.7	11.0
LLAVA-OV [37] (Qwen2 7B)	55.5	12.6	4.9	0.0	14.2	5.0	0.1	0.0	6.2	4.0	10.2
PHI-3-VISION [1]	53.4	10.9	0.8	0.4	12.0	21.6	6.5	**0.1**	14.7	6.5	12.7
QWEN2VL [60] 2B	60.8	12.1	0.4	**25.6**	**42.9**	48.5	15.7	**0.1**	29.0	10.8	**24.6**
QWEN2VL [60] 7B	**63.2**	**15.7**	2.7	1.4	42.3	**49.3**	12.1	**0.1**	**29.5**	12.5	22.9
Open-world baselines											
CaSED [17]	35.5	5.1	3.0	1.4	28.1	19.4	**34.6**	0.0	13.5	8.1	14.9
CLIP retrieval	42.6	7.5	6.6	14.0	40.6	26.4	30.3	0.0	14.7	8.4	19.1
Closed-world baselines											
CLIP [49]	87.1	52.6	42.7	27.2	76.9	89.9	88.1	76.2	65.6	72.7	67.9
SigLIP [68]	93.6	60.8	42.1	46.0	88.2	94.1	95.4	92.3	69.9	82.1	76.5

Table 5. Text inclusion on the ten datasets. Higher is better, **bold** indicates best.

Model	Llama inclusion										
	C101	DTD	ESAT	FGVC	FLWR	FOOD	PETS	CARS	S397	U101	Avg.
IDEFICS2 [34] 8B	72.9	24.6	19.0	64.4	54.6	58.7	36.3	69.6	32.5	40.1	47.3
INSTRUCTBLIP [19] Vicuna 7B	76.8	26.2	19.1	59.9	57.4	47.6	41.3	62.0	35.8	36.0	46.2
INTERNVL2 [12, 13] 2B	74.9	48.5	35.0	35.8	49.3	44.3	47.4	30.0	64.9	52.1	48.2
INTERNVL2 [12, 13] 4B	74.4	45.7	30.1	40.5	37.5	45.9	49.7	33.1	62.5	50.4	47.0
INTERNVL2 [12, 13] 8B	77.2	50.5	28.6	29.7	36.0	53.7	50.4	35.3	71.5	**59.6**	49.3
LLAVA-1.5 [41] 7B	74.5	39.4	**45.0**	44.5	46.3	47.7	45.5	37.5	51.6	48.5	48.1
LLAVA-NEXT [36] (Mistral 7B)	77.8	54.0	28.0	43.4	33.4	63.2	34.6	50.9	69.9	58.3	51.4
LLAVA-NEXT [36] (Vicuna 7B)	77.3	52.2	26.4	43.1	29.2	60.6	43.6	41.2	68.2	59.1	50.1
LLAVA-OV [37] (Qwen2 0.5B)	76.5	46.5	28.7	61.2	55.1	28.1	44.9	70.0	52.2	35.8	49.9
LLAVA-OV [37] (Qwen2 7B)	81.3	45.6	11.8	**68.9**	48.9	22.0	50.2	**84.4**	25.0	27.0	46.5
PHI-3-VISION [1]	75.7	45.3	6.0	51.0	53.2	45.1	49.1	39.0	44.5	34.7	44.4
QWEN2VL [60] 2B	82.9	54.6	3.1	65.0	67.0	71.1	49.3	56.3	72.6	45.2	56.7
QWEN2VL [60] 7B	**84.3**	**60.8**	18.1	58.8	**71.0**	**75.0**	46.0	67.2	**73.0**	48.8	**60.3**
Open-world baselines											
CaSED [17]	57.7	16.7	7.3	30.7	46.0	35.1	**58.7**	63.5	34.9	31.7	38.2
CLIP retrieval	55.3	28.2	12.7	25.8	44.6	35.4	56.2	10.4	30.5	32.9	33.2
Closed-world baselines											
CLIP [49]	87.1	52.6	42.7	27.2	76.9	89.9	88.1	76.2	65.6	72.7	67.9
SigLIP [68]	93.6	60.8	42.1	46.0	88.2	94.1	95.4	92.3	69.9	82.1	76.5

Table 6. Llama inclusion on the ten datasets. Higher is better, **bold** indicates best. Note that the scores for CLIP closed-world equals the textual inclusion scores.

open-source models (*i.e.*, 13 base + 5 reasoning and the larger models from the previous analysis). The estimated sizes of these models are 8B (Haiku and GPT-4o-mini), 32B (Gemini 2.0 Flash), and +175B (Sonnet and GPT-4o). From the results, we notice there isn't a large gap between open and commercial models, with GPTs and Gemini performing on par with, *e.g.*, InternVL2 26B and Qwen2-VL 72B. Only Claude consistently achieves better performance, still com-

Model	Semantic similarity										
	C101	DTD	ESAT	FGVC	FLWR	FOOD	PETS	CARS	S397	U101	Avg.
IDEFICS2 [34] 8B	64.9	34.6	27.5	27.6	38.6	44.4	30.8	31.6	44.2	44.0	38.8
INSTRUCTBLIP [19] Vicuna 7B	**71.5**	32.8	30.0	21.4	38.9	41.6	26.4	38.5	42.1	48.3	39.1
INTERNVL2 [12, 13] 2B	50.5	25.6	26.0	23.4	31.2	39.6	23.9	42.9	43.3	43.1	34.9
INTERNVL2 [12, 13] 4B	49.2	26.1	24.7	23.6	30.2	41.1	24.6	44.1	43.8	41.8	34.9
INTERNVL2 [12, 13] 8B	50.1	26.7	24.4	25.5	32.8	44.2	27.3	46.6	46.3	44.6	36.8
LLAVA-1.5 [41] 7B	49.0	24.2	**34.2**	19.0	25.8	37.2	21.5	38.2	41.7	40.7	33.1
LLAVA-NEXT [36] (Mistral 7B)	48.2	27.7	23.9	23.6	30.2	45.3	30.3	44.8	43.6	42.1	36.0
LLAVA-NEXT [36] (Vicuna 7B)	49.2	27.9	23.1	23.4	29.3	43.0	24.4	45.7	43.3	42.3	35.1
LLAVA-OV [37] (Qwen2 0.5B)	64.7	28.8	21.6	21.0	41.4	42.7	31.4	40.0	43.2	47.9	38.3
LLAVA-OV [37] (Qwen2 7B)	68.7	32.2	19.4	29.4	37.5	41.7	37.8	34.4	43.4	43.2	38.8
PHI-3-VISION [1]	53.6	28.5	12.3	18.8	30.9	40.1	24.3	39.0	41.8	37.3	32.7
QWEN2VL [60] 2B	56.4	27.0	13.5	**32.8**	43.7	50.6	27.8	**57.4**	47.9	42.7	40.0
QWEN2VL [60] 7B	55.8	28.5	20.7	20.6	41.8	50.6	25.1	48.5	48.1	43.2	38.3
Open-world baselines											
CaSED [17]	65.3	**39.9**	32.2	30.0	**55.6**	**64.1**	**62.4**	47.1	**52.4**	**53.4**	**50.2**
CLIP retrieval	41.3	23.6	22.4	30.7	40.3	46.7	41.7	48.8	39.1	38.5	37.3
Closed-world baselines											
CLIP [49]	90.8	69.9	67.7	66.7	83.4	93.7	91.8	80.5	92.2	83.3	82.0
SigLIP [68]	97.8	75.6	63.1	80.0	92.0	96.4	96.8	98.1	83.1	89.6	87.3

Table 7. Semantic similarity on ten datasets. Higher is better, **bold** indicates best.

Model	Concept similarity										
	C101	DTD	ESAT	FGVC	FLWR	FOOD	PETS	CARS	S397	U101	Avg.
IDEFICS2 [34] 8B	76.3	38.5	30.9	29.7	41.5	48.4	35.3	37.5	49.9	54.6	44.3
INSTRUCTBLIP [19] Vicuna 7B	75.3	39.1	31.6	28.6	43.6	60.0	37.9	40.0	52.6	55.3	46.4
INTERNVL2 [12, 13] 2B	75.7	48.0	**52.9**	36.8	49.5	60.8	41.9	50.9	65.1	59.4	54.1
INTERNVL2 [12, 13] 4B	76.1	48.6	51.5	37.9	51.0	63.0	41.9	50.5	65.4	59.1	54.5
INTERNVL2 [12, 13] 8B	78.7	49.7	49.1	42.5	56.9	67.1	46.0	56.2	69.2	**62.9**	57.8
LLAVA-1.5 [41] 7B	72.1	41.3	51.6	29.0	41.6	56.8	35.9	46.2	59.4	55.5	48.9
LLAVA-NEXT [36] (Mistral 7B)	79.8	**51.0**	49.5	37.5	55.1	70.0	55.3	56.3	68.7	62.7	58.6
LLAVA-NEXT [36] (Vicuna 7B)	79.0	50.1	50.8	37.1	51.3	65.8	42.4	55.0	67.4	61.8	56.1
LLAVA-OV [37] (Qwen2 0.5B)	77.8	45.1	39.9	30.6	42.4	50.0	37.5	43.5	56.7	55.9	47.9
LLAVA-OV [37] (Qwen2 7B)	79.1	47.0	41.0	29.4	51.7	41.9	37.8	35.4	44.9	43.3	45.1
PHI-3-VISION [1]	74.1	44.0	25.3	29.1	43.0	58.3	40.3	42.9	56.1	49.1	46.2
QWEN2VL [60] 2B	79.4	47.3	24.2	**56.0**	67.9	75.7	46.7	**68.6**	70.0	56.6	**59.2**
QWEN2VL [60] 7B	**81.3**	50.4	39.8	30.8	**68.8**	**76.9**	43.1	56.0	**70.6**	59.1	57.7
Open-world baselines											
CaSED [17]	65.9	39.8	32.2	29.9	55.6	66.5	62.9	47.1	53.7	55.1	50.9
CLIP retrieval	63.9	38.1	37.8	50.7	62.3	67.8	**66.1**	61.5	57.3	54.4	56.0
Closed-world baselines											
CLIP [49]	90.8	69.9	67.7	66.7	83.4	93.7	91.8	80.5	92.2	83.3	82.0
SigLIP [68]	97.8	75.6	63.1	80.0	92.0	96.4	96.8	98.1	83.1	89.6	87.3

Table 8. Concept similarity on ten datasets. Higher is better, **bold** indicates best.

parable to Qwen2.5-VL 7B at a fraction of the size. Surprisingly, GPT-4o-mini is better than GPT-4o on the task, similarly to the findings for InternVL2 2B *vs*. 4B. Also, *reasoning models are strong*: Qwen2.5-VL 7B outperforms Qwen2-VL 72B and the commercial models on most of the metrics despite its reduced dimension.

Linking model info with performance. Many models do not disclose their full training details, making it hard to identify key factors influencing performance. However, by linking the results in Tab. 1 with the summary in Tab. 4,

Rank	Average Elo ratings Model	Rating
1	QWEN2VL [60] 2B	1037
2	QWEN2VL [60] 7B	1037
3	PHI-3-VISION [1]	1029
4	LLAVA-NEXT [36] (Mistral 7B)	1018
5	LLAVA-NEXT [36] (Vicuna 7B)	1015
6	LLAVA-OV [37] (Qwen2 7B)	1014
7	LLAVA-OV [37] (Qwen2 0.5B)	1007
8	INTERNVL2 [12, 13] 8B	1004
9	INTERNVL2 [12, 13] 4B	994
10	INTERNVL2 [12, 13] 2B	991
11	LLAVA-1.5 [41] 7B	984
12	INSTRUCTBLIP [19] Vicuna 7B	943
13	IDEFICS2 [34] 8B	924

Table 9. Elo ratings on the ten datasets. Higher scores indicate comparatively better responses from the models.

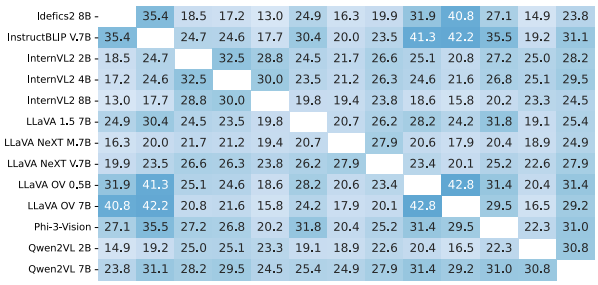


Figure 9. Percentage of correct but generic predictions shared between models. Higher values indicate models perform responses similarly to the same inputs.

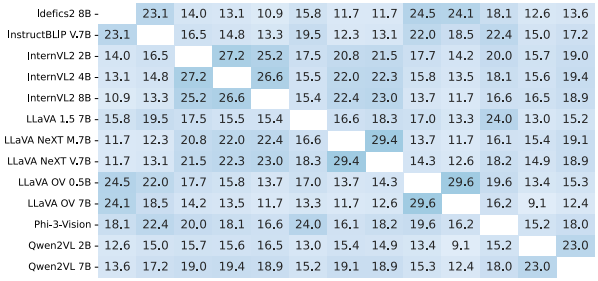


Figure 10. Percentage of wrong but specific predictions shared between models. Higher values indicate models perform responses similarly to the same inputs.

we hypothesize that (i) the pre-training of the vision encoder is more important than the size (*e.g.*, InternVL and Qwen2-VL *vs*. CLIP/SigLIP, or *vs*. BLIP-2, which has double the size); (ii) higher image resolution can improve performance (*e.g.*, LLaVA-NeXT *vs*. LLaVA-1.5); (iii) the pre-training of the language encoder is less important than the training strategy (*e.g.*, LLaVA-NeXT Mistral/Vicuna *vs*. Idefics/InstructBLIP); (iv) the size of the language encoder is not an indicator of performance (*e.g.*, LLaVA-NeXT Mistral *vs*. Yi, GPT-4o-mini *vs*. GPT-4o). It is also reasonable to assume that the strongest influence comes from the training data for which details are only partly available.

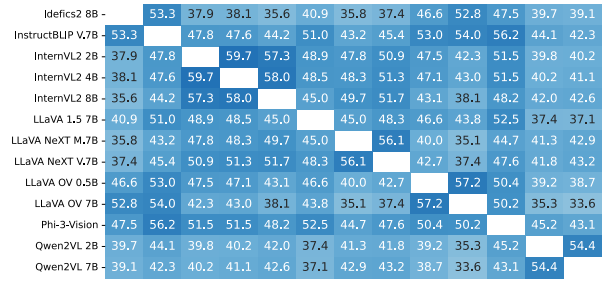


Figure 11. Percentage of wrong and generic predictions shared between models. Higher values indicate models perform responses similarly to the same inputs.

A.5. Extended results for the analyses

Below, we report the extended results for the analyses we conducted. In Tab. 15 (also visualizing the average gains in Fig. 12) we show the variation in correct and wrong predictions for each model when using more generic/specific prompts and domain-specific information. We additionally report the variation in text inclusion, Llama inclusion, and concept similarity for each model and dataset in Tab. 13 and Tab. 14. For the chain-of-thought experiments, we provide the variations on the correct and wrong predictions in Tab. 17, and the per-dataset and model variations in Tab. 16. We also provide a visualization of the variations for the list, caption, and describe experiments in Fig. 13 (also reported numerically in Tab. 18). Finally, we report the complete results table for the reasoning models tested on the ten classification datasets in Tab. 19.

Model	Prototypical				Non-prototypical				Fine-grained				Very fine-grained			
	TI	LI	SS	CS	TI	LI	SS	CS	TI	LI	SS	CS	TI	LI	SS	CS
IDEFICS2 [34] 8B	30.8	52.7	54.5	63.1	3.7	27.9	35.4	41.3	3.0	49.9	38.0	41.7	0.0	67.0	29.6	33.6
INSTRUCTBLIP [19] Vicuna 7B	29.7	56.3	56.8	64.0	6.0	27.1	37.0	42.0	10.4	48.8	35.6	47.2	0.0	61.0	30.0	34.3
(*) INSTRUCTBLIP [19] Vicuna 13B	22.7	47.7	49.5	57.8	4.4	27.3	34.2	41.2	6.6	36.7	30.2	41.4	0.0	52.9	31.5	34.2
INTERNVL2 [12, 13] 2B	36.9	69.9	46.9	70.4	10.2	45.2	31.6	53.4	14.9	47.0	31.6	50.7	0.7	32.9	33.1	43.9
INTERNVL2 [12, 13] 4B	36.3	68.5	46.5	70.8	10.1	42.1	30.8	53.1	16.2	44.4	32.0	52.0	1.7	36.8	33.8	44.2
INTERNVL2 [12, 13] 8B	40.6	74.4	48.2	74.0	11.0	46.2	31.9	53.9	22.3	46.7	34.8	56.7	2.3	32.5	36.0	49.4
(*) INTERNVL2 [12, 13] 26B	46.6	78.6	49.1	**77.7**	**15.8**	**58.7**	36.7	**60.5**	36.5	58.9	40.2	65.0	7.1	40.8	40.9	59.3
LLAVA-1.5 [41] 7B	34.6	63.1	45.3	65.8	8.6	44.3	33.0	49.5	8.4	46.5	28.2	44.8	0.0	41.0	28.6	37.6
(*) LLAVA-1.5 [41] 13B	35.7	63.5	47.0	66.7	9.5	43.0	34.1	51.2	8.8	48.0	28.7	44.9	0.0	37.4	28.9	37.8
LLAVA-NEXT [36] (Mistral 7B)	41.7	73.9	45.9	74.3	11.3	46.8	31.2	54.4	26.8	43.7	35.3	60.1	1.4	47.2	34.2	46.9
LLAVA-NEXT [36] (Vicuna 7B)	39.5	72.8	46.2	73.2	10.6	45.9	31.1	54.2	16.9	44.5	32.2	53.2	1.3	42.2	34.5	46.1
(*) LLAVA-NEXT [36] (Vicuna 13B)	42.2	73.6	46.2	75.3	11.4	46.5	32.4	55.5	26.2	44.0	36.1	60.4	1.3	33.4	34.4	47.0
(*) LLAVA-NEXT [36] (Yi 34B)	39.2	74.9	46.2	73.9	12.2	49.3	33.1	56.6	25.3	43.1	35.1	60.0	0.9	42.5	33.5	45.3
LLAVA-OV [37] (Qwen2 0.5B)	34.4	64.4	54.0	67.3	7.3	37.0	32.8	47.0	6.0	42.7	38.5	43.3	0.6	65.6	30.5	37.1
LLAVA-OV [37] (Qwen2 7B)	30.8	53.2	56.1	62.0	7.2	28.1	31.6	43.8	6.4	40.4	39.0	43.8	0.0	76.7	31.9	32.4
PHI-3-VISION [1]	34.1	60.1	47.7	65.1	6.0	28.7	26.0	39.5	13.4	49.1	31.8	47.2	0.2	45.0	28.9	36.0
QWEN2VL [60] 2B	44.9	77.8	52.2	74.7	7.8	34.3	27.7	42.7	35.7	62.5	40.7	63.4	**12.9**	60.7	**45.1**	**62.3**
QWEN2VL [60] 7B	46.4	**78.7**	51.9	76.0	10.3	42.6	30.8	49.8	34.6	64.0	39.2	62.9	0.8	**63.0**	34.5	43.4
(*) QWEN2VL [60] 72B	**47.7**	78.2	49.1	76.7	10.4	42.1	28.6	48.6	**48.1**	**66.6**	43.4	**71.8**	11.9	59.1	40.6	58.8
Open-world baselines																
CASED [17]	24.5	46.3	**58.9**	59.8	5.4	18.6	**41.8**	42.4	27.4	46.6	**60.7**	61.7	0.7	47.1	38.5	38.5
CLIP retrieval	28.6	42.9	40.2	60.6	7.5	24.6	28.1	43.4	32.4	45.4	42.9	65.4	7.0	18.1	39.7	56.1
Closed-world baselines																
CLIP [49]	76.4		91.5		56.0		73.6		85.0		89.6		51.7		73.6	
SigLIP [68]	81.8		90.5		61.7		76.1		92.6		95.1		69.2		89.1	

Table 10. OW results with larger models (in green) averaged on the grouped datasets. TI stands for text inclusion, LI for Llama inclusion, SS for semantic similarity, and CS for concept similarity. Higher is better, **bold** indicates best.

Model	TI	LI	SS	CS
IDEFICS2 8B	12.5	45.7	42.6	49.2
INSTRUCTBLIP Vicuna 7B	13.4	38.3	37.4	50.2
(*) INSTRUCTBLIP Vicuna 13B	10.0	38.3	37.4	45.2
INTERNVL2 2B	19.5	54.4	34.9	54.9
INTERNVL2 4B	18.9	51.5	34.4	55.3
INTERNVL2 8B	22.9	54.7	36.3	58.8
(*) INTERNVL2 26B	31.3	63.1	39.5	63.9
LLAVA-1.5 7B	14.7	50.8	32.2	49.3
(*) LLAVA-1.5 13B	15.7	52.9	33.1	50.1
LLAVA-NEXT (Mistral 7B)	25.9	51.6	35.7	60.8
LLAVA-NEXT (Vicuna 7B)	20.2	52.3	34.6	56.9
(*) LLAVA-NEXT (Vicuna 13B)	25.7	52.6	36.4	61.4
(*) LLAVA-NEXT (Yi 34B)	24.5	52.4	36.1	60.7
LLAVA-OV (Qwen2 0.5B)	15.3	51.8	42.8	51.7
LLAVA-OV (Qwen2 7B)	17.3	50.6	**43.9**	51.8
PHI-3-VISION	17.9	51.6	34.9	50.1
QWEN2VL 2B	28.5	59.8	39.5	59.6
QWEN2VL 7B	29.2	62.2	38.9	60.5
(*) QWEN2VL 72B	36.7	64.0	40.2	66.3

Model	TI	LI	SS	CS
Reasoning models				
INTERNVL2.5 2B	20.3	51.7	33.1	54.6
INTERNVL2.5 4B	21.6	54.4	35.7	55.8
INTERNVL2.5 8B	21.3	55.9	36.0	56.3
QWEN2.5VL 3B	36.5	64.2	39.8	66.1
QWEN2.5VL 7B	**45.0**	71.5	41.9	**72.6**
Commercial models				
(*) GPT-4O-MINI	29.5	70.3	39.9	63.1
(*) GPT-4O	27.4	66.3	41.2	59.9
(*) CLAUDE HAIKU 3.5	37.2	74.7	42.1	70.1
(*) CLAUDE SONNET 3.5	39.0	**77.3**	42.4	72.2
(*) GEMINI 2.0 FLASH	29.2	62.2	39.1	60.1
Open-world baselines				
CASED	22.3	42.2	55.3	55.9
CLIP retrieval	25.9	43.4	37.0	57.0
Closed-world baselines				
CLIP	75.5		83.8	
SigLIP	84.0		90.4	

Table 11. Results with larger (in green) and commercial (in purple) models averaged on 5 datasets, *i.e.*, Caltech101, DTD, Flowers102, OxfordPets, UCF101. TI stands for text inclusion, LI for Llama inclusion, SS for semantic similarity, and CS for concept similarity. Higher is better, **bold** is best.

Model	Caltech101			DTD			Flowers102			OxfordPets			UCF101		
	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS
List															
IDEFICS2 [34] 8B	-1.5	-12.8	-2.8	+3.6	+3.4	+2.0	+3.6	-19.3	-0.3	+1.8	-17.6	+3.5	+2.7	-6.6	-1.2
INSTRUCTBLIP [19] Vicuna 7B	+12.5	-3.5	+4.5	+6.8	+15.8	+6.0	+13.7	-24.2	+5.9	+2.6	-29.1	+0.9	+6.7	+21.1	+9.3
INTERNVL2 [12, 13] 2B	-0.8	-6.1	+1.2	-1.5	-2.8	-1.7	-4.2	-23.5	-2.4	+4.8	-32.8	+2.2	-3.1	-11.2	-1.6
INTERNVL2 [12, 13] 4B	+1.8	-4.3	+0.7	-1.8	+3.6	-1.6	-3.7	-12.7	-5.7	+3.4	-35.5	+3.1	+2.2	-0.2	+0.6
INTERNVL2 [12, 13] 8B	-1.7	-2.2	-0.2	-0.8	+0.1	-1.2	-4.8	-2.3	-5.2	+2.3	-26.5	+2.4	+1.0	-3.9	-3.2
LLAVA-1.5 [41] 7B	-0.7	-6.3	-2.7	+0.7	+2.6	+1.4	+0.8	-21.5	-0.9	+0.9	-23.0	+1.4	-1.5	-10.2	+0.0
LLAVA-NEXT [36] (Mistral 7B)	-2.5	-2.1	-0.4	-1.1	-1.2	-0.8	-5.1	-2.1	-4.1	-9.5	-7.6	-4.3	-0.6	-0.6	-1.8
LLAVA-NEXT [36] (Vicuna 7B)	-1.5	-2.9	-0.6	-0.7	-1.4	-1.0	-2.7	-3.7	-2.8	+3.6	-19.0	+6.5	-1.3	-3.7	-1.5
LLAVA-OV [37] (Qwen2 0.5B)	+6.3	-8.1	+0.5	+1.8	+5.7	+4.4	+8.1	-23.8	+4.0	+2.1	-39.0	+3.0	+5.5	+19.5	+4.0
LLAVA-OV [37] (Qwen2 7B)	+6.5	+1.1	+3.6	+3.0	+14.2	+6.2	+1.6	-7.2	-0.6	+2.1	-46.9	+0.9	+13.6	+29.1	+19.9
PHI-3-VISION [1]	-1.1	-11.9	+1.0	-1.9	+1.6	-1.8	+3.3	-17.1	+1.4	+1.5	-40.7	+1.9	+1.2	+3.0	+3.9
QWEN2VL [60] 2B	+0.4	-0.1	+2.6	+5.9	+9.7	+6.3	-10.6	-12.2	-8.0	-3.4	-23.1	-5.0	+6.3	+18.6	+7.5
QWEN2VL [60] 7B	-1.8	-10.4	-1.0	+0.4	+2.7	+1.4	-24.2	-44.0	-21.4	-5.6	-39.2	-1.0	+0.4	+0.8	-0.7
Caption															
IDEFICS2 [34] 8B	-2.8	-5.4	-4.9	+5.3	+13.5	+3.2	+6.2	-33.5	-0.7	+8.3	-18.9	+6.4	+3.4	+1.3	+0.2
INSTRUCTBLIP [19] Vicuna 7B	+9.4	-8.6	+0.0	+4.7	+14.6	+3.4	+6.6	-34.5	-0.8	+3.5	-24.6	+1.6	+5.2	+11.1	+4.5
INTERNVL2 [12, 13] 2B	-6.3	-4.5	-9.9	-1.1	-5.1	-6.3	-2.2	-20.3	-7.3	+10.3	-13.3	+3.0	+0.2	-7.0	-2.4
INTERNVL2 [12, 13] 4B	-0.8	-6.5	-9.2	-2.2	-2.9	-5.3	-1.9	-13.6	-7.9	+13.8	-18.3	+4.3	-0.1	-3.6	-19.7
INTERNVL2 [12, 13] 8B	-4.2	-4.3	-9.5	+1.0	+0.6	-4.8	-5.4	-4.2	-13.5	+2.9	-20.3	-1.5	-1.4	-10.0	-4.6
LLAVA-1.5 [41] 7B	+0.9	+1.6	+5.4	+3.0	+4.8	+7.1	-1.0	-27.5	+3.7	+5.6	-19.6	+7.9	+3.2	+9.2	+5.3
LLAVA-NEXT [36] (Mistral 7B)	-9.4	-15.5	-17.7	-5.5	-20.9	-9.0	-8.1	-10.2	-15.2	-16.2	-10.6	-17.8	-4.8	-28.0	-9.7
LLAVA-NEXT [36] (Vicuna 7B)	-8.4	-13.5	-16.9	-4.9	-20.3	-9.7	-2.2	-5.9	-11.3	+0.5	-17.0	-3.8	-4.8	-24.3	-9.5
LLAVA-OV [37] (Qwen2 0.5B)	+2.5	-0.1	-8.9	+1.2	+2.7	+0.2	+12.6	-23.0	-0.3	+10.2	-18.2	+4.7	+7.1	+22.6	+5.2
LLAVA-OV [37] (Qwen2 7B)	+0.7	-6.8	-8.1	-1.7	+1.9	-3.1	-6.2	-28.4	-11.2	+5.6	-35.4	+2.0	+9.5	+21.3	+16.9
PHI-3-VISION [1]	+2.5	+0.4	+3.5	+2.4	+10.0	+5.8	+12.4	-4.6	+8.7	+17.1	-14.2	+12.8	+3.8	+13.5	+4.9
QWEN2VL [60] 2B	+0.3	+0.3	-5.2	+2.9	+7.1	+0.5	-6.6	-8.9	-14.8	+20.0	+0.2	+8.5	+3.6	+15.4	+5.2
QWEN2VL [60] 7B	-0.6	-1.1	+0.9	+1.7	+1.7	+1.2	-8.5	-15.8	-13.5	+13.2	-7.2	+10.0	+48.4	-32.7	+3.8
Describe															
IDEFICS2 [34] 8B	-10.0	-21.1	-5.1	-0.8	-4.9	-2.2	+3.3	-33.0	-0.6	+2.1	-22.0	+4.0	-2.4	-19.9	-6.8
INSTRUCTBLIP [19] Vicuna 7B	+11.5	-4.6	-1.9	+6.9	+18.7	+5.2	+15.2	-25.5	+3.5	+9.6	-12.3	+6.8	+5.0	+24.1	+8.3
INTERNVL2 [12, 13] 2B	-1.4	+1.4	+3.4	+0.9	-0.5	+1.0	-2.3	-18.0	+1.1	+14.9	-18.3	+11.4	+9.7	-23.0	-6.1
INTERNVL2 [12, 13] 4B	+2.0	-0.9	+2.2	+34.6	+5.4	-20.8	-0.4	-10.6	+0.1	+16.3	-20.4	+12.4	+1.7	+1.6	+0.8
INTERNVL2 [12, 13] 8B	-1.7	-0.2	+1.0	+0.3	+2.1	+0.4	-5.1	+0.0	-5.4	+10.7	-18.7	+9.0	+0.9	-1.6	-2.1
LLAVA-1.5 [41] 7B	+1.9	+0.6	+6.6	+3.8	+6.6	+7.5	+0.5	-28.0	+4.7	+5.6	-22.9	+7.9	+2.7	+10.0	+5.7
LLAVA-NEXT [36] (Mistral 7B)	-1.9	+0.8	+0.8	+0.3	-1.0	+0.8	-4.9	-3.8	-4.1	-6.6	-3.1	-1.9	+1.1	+3.6	-1.1
LLAVA-NEXT [36] (Vicuna 7B)	-1.7	+0.5	+0.4	-0.5	-0.7	+0.4	-2.7	-5.1	-3.7	+7.6	-13.1	+9.5	-1.1	+1.6	-1.8
LLAVA-OV [37] (Qwen2 0.5B)	+6.7	+7.6	+4.3	+6.1	+21.7	+8.0	+12.1	-6.9	+7.6	+11.7	-7.6	+10.7	+5.4	+1.5	-7.7
LLAVA-OV [37] (Qwen2 7B)	+8.1	+6.5	+5.8	+5.1	+22.4	+7.0	+7.4	-5.3	+3.0	+20.3	-18.0	+14.3	+16.6	+43.1	+21.6
PHI-3-VISION [1]	-0.1	-0.1	+2.0	+2.8	+11.5	+5.6	+11.7	-2.7	+8.5	+14.6	-14.0	+11.2	+4.4	+15.0	+5.8
QWEN2VL [60] 2B	+1.8	+1.8	+5.0	+5.4	+8.1	+6.7	-6.5	-13.3	-3.8	+12.8	-12.1	+10.0	+17.7	-8.0	+0.1
QWEN2VL [60] 7B	-1.0	-1.9	+2.9	+2.1	+2.4	+4.1	-2.6	-15.6	-4.1	+24.6	-2.2	+18.6	+4.4	+14.2	+4.5

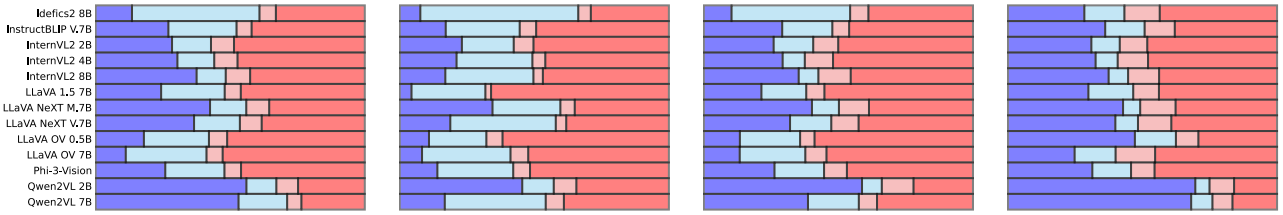
Table 12. Relative performance variation with multi-label prompts on five datasets. TI stands for text inclusion, LI for Llama inclusion, and CS for concept similarity.

Model	DTD			FGVCAircraft			Flowers102			Food101			OxfordPets			StanfordCars		
	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS
Be generic																		
IDEFICS2 [34] 8B	-7.0	-9.7	-8.5	-2.4	+27.2	-2.4	-25.7	+8.9	-13.3	-19.8	+30.5	-20.7	-0.5	-15.5	-4.2	+0.0	+8.2	-19.1
INSTRUCTBLIP [19] Vicuna 7B	-10.6	-20.7	-9.3	-0.7	-9.1	-4.4	-29.6	+2.5	-19.0	-15.7	-13.9	-13.3	-13.6	-20.5	-8.4	+0.0	+10.2	-24.5
INTERNVL2 [12, 13] 2B	-3.0	+3.8	-3.9	-2.5	+43.5	-15.8	-6.1	+31.8	-11.4	-4.5	-11.5	-6.2	-14.8	-8.2	-9.1	+0.0	+13.9	-8.2
INTERNVL2 [12, 13] 4B	-6.3	-4.7	-3.9	-7.2	+63.2	-20.1	-13.7	+35.8	-15.2	-13.3	-11.6	-10.4	-10.0	-3.6	-7.1	-0.0	+51.4	-14.1
INTERNVL2 [12, 13] 8B	-12.4	-16.1	-10.8	-7.3	+44.3	-21.3	-20.1	+37.7	-17.9	-21.3	-18.0	-17.9	-15.0	-3.0	-9.6	-0.0	+65.6	-18.3
LLAVA-1.5 [41] 7B	-4.4	-21.7	-7.0	-0.2	+23.1	-0.3	-10.5	+27.0	-2.2	-23.0	-41.7	-27.2	-5.0	-20.7	-4.1	+0.0	+45.3	-19.9
LLAVA-NEXT [36] (Mistral 7B)	-4.5	-6.0	-3.5	-5.9	+27.6	-17.7	-13.4	+25.5	-13.1	-8.5	-0.4	-5.2	-19.6	-12.3	-14.2	-0.0	+43.2	-21.7
LLAVA-NEXT [36] (Vicuna 7B)	-3.1	-9.7	-5.6	-5.3	+59.6	-17.3	-12.8	+37.2	-12.3	-19.0	-10.2	-16.2	-23.9	-0.7	-19.2	+0.0	+50.1	-24.5
LLAVA-OV [37] (Qwen2 0.5B)	-10.1	-20.4	-6.9	-8.7	+5.3	-18.2	-15.9	+18.2	-16.3	-30.2	-51.5	-26.5	-25.6	-29.4	-20.3	-0.1	+3.4	-27.6
LLAVA-OV [37] (Qwen2 7B)	-39.1	-32.8	-26.1	-2.4	+4.8	-7.7	-29.8	+4.9	-21.4	-11.2	-2.3	-15.0	-0.0	-0.7	-0.3	+0.0	+4.8	-24.4
PHI-3-VISION [1]	-5.7	-0.1	+0.5	-1.8	+8.1	-5.0	-28.0	-2.1	-13.5	-14.7	-17.8	-11.8	-16.9	-30.2	-9.9	+0.0	+39.4	-15.9
QWEN2VL [60] 2B	-8.0	-13.4	-8.9	-35.5	-1.4	-36.6	-16.3	+1.7	-17.8	-18.4	-22.5	-13.5	-41.6	-22.8	-26.9	-0.0	-18.4	-6.3
QWEN2VL [60] 7B	-11.1	-17.6	-10.5	-39.6	-9.9	-43.6	-55.7	-20.8	-40.0	-41.5	-31.4	-30.2	-15.4	-11.1	-10.0	+0.0	-4.3	-34.4
Be specific																		
IDEFICS2 [34] 8B	-1.0	-4.2	-1.7	+0.0	-3.5	-0.2	+0.2	-2.3	+0.0	-2.2	-5.6	-2.3	-0.1	-5.4	-1.5	+0.0	+5.4	-1.8
INSTRUCTBLIP [19] Vicuna 7B	+1.9	+13.9	+0.1	+0.0	-12.4	-1.8	+5.5	+5.6	+1.8	-0.2	+2.9	-0.3	-0.1	-4.7	-0.9	+0.0	-22.3	+7.1
INTERNVL2 [12, 13] 2B	-0.5	-4.5	-0.7	-0.8	+5.9	-3.6	-2.2	+8.1	-1.8	+0.1	-3.4	-0.5	-2.4	-1.8	-1.2	-0.0	+1.2	-2.6
INTERNVL2 [12, 13] 4B	-0.8	-0.1	-1.1	+1.2	-11.1	+2.7	+0.9	-3.5	-1.3	+0.1	-3.3	-1.3	+7.6	-4.9	+3.6	-0.0	-15.1	+2.1
INTERNVL2 [12, 13] 8B	+1.5	+0.0	+0.2	+0.1	-6.2	+1.5	-0.5	-4.5	+0.2	-0.2	-3.2	-0.5	+2.0	-7.1	+1.4	+0.0	-13.4	+1.6
LLAVA-1.5 [41] 7B	+0.5	+0.3	-0.7	+0.0	-4.1	-0.6	+0.6	-4.1	-0.6	-0.3	-8.1	-2.2	+0.3	+0.9	-0.8	+0.0	-11.7	-0.4
LLAVA-NEXT [36] (Mistral 7B)	+0.7	-2.0	+0.2	+1.1	-12.6	+2.5	+0.2	-3.0	-0.2	-0.5	-2.6	-0.4	-0.9	-1.8	-0.5	+0.0	-18.1	+0.3
LLAVA-NEXT [36] (Vicuna 7B)	-0.8	-3.1	-1.0	-0.6	-5.8	-3.5	-3.0	+16.9	-4.0	-1.5	-5.8	-1.6	-3.4	-2.7	-3.0	-0.0	-14.5	-1.6
LLAVA-OV [37] (Qwen2 0.5B)	-1.3	-10.0	-3.6	-1.2	-6.3	-1.4	+4.2	+7.2	+3.0	-3.6	-6.0	-4.5	+0.5	-1.2	+0.8	-0.1	-2.4	-7.2
LLAVA-OV [37] (Qwen2 7B)	+1.1	+5.5	+1.5	+0.0	-3.9	+0.0	+2.6	+0.7	+1.4	+2.6	-0.8	+3.2	+0.3	-4.8	+0.3	+0.0	-5.5	+2.9
PHI-3-VISION [1]	+0.7	+3.6	+0.4	+0.3	+6.2	+0.3	+5.2	+0.3	+4.1	+3.4	-0.9	+1.0	+2.4	-1.6	+1.4	-0.1	-2.5	+1.6
QWEN2VL [60] 2B	+4.0	+4.1	+3.1	+6.4	-5.7	+9.1	+4.1	-3.2	+4.6	+2.9	+1.3	+1.9	+13.4	+5.5	+7.8	+0.1	-3.4	+4.5
QWEN2VL [60] 7B	-2.8	-8.1	-4.1	-1.0	+0.8	-1.0	-21.9	-15.0	-16.3	-11.2	-17.8	-8.7	-9.1	-3.9	-4.7	+0.0	-0.3	-5.5

Table 13. Relative performance variation with the generic/specific prompts on six datasets. TI stands for text inclusion, LI for Llama inclusion, and CS for concept similarity.

Model	DTD			FGVCAircraft			Flowers102			Food101			OxfordPets			StanfordCars		
	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS
IDEFICS2 [34] 8B	+6.0	+3.2	+6.8	+2.4	-37.3	+2.2	+24.9	-10.2	+12.9	+15.3	-19.6	+15.0	+0.4	+9.9	+3.2	+0.0	-2.1	+16.8
INSTRUCTBLIP [19] Vicuna 7B	+9.0	+11.2	+6.1	+0.7	-0.3	+4.1	+26.6	-11.1	+17.9	+6.4	+1.5	+5.7	+12.8	+16.6	+7.9	+0.0	-7.4	+19.9
INTERNVL2 [12, 13] 2B	+2.3	-5.9	+0.6	+1.5	-20.9	+7.6	+2.2	-24.6	+7.1	+0.8	+5.4	+1.8	+9.8	+5.9	+6.5	-0.0	-8.4	+3.1
INTERNVL2 [12, 13] 4B	+3.8	+0.9	+1.3	+3.8	-25.3	+10.2	+5.8	-12.0	+5.6	+3.0	+4.1	+2.2	+4.7	+3.6	+3.8	+0.0	-18.5	+4.9
INTERNVL2 [12, 13] 8B	+6.5	+1.5	+1.5	+2.6	-13.9	+8.4	+1.9	-9.4	+2.4	+2.0	+0.5	+1.5	+2.0	+3.8	+2.0	-0.1	-14.9	+0.8
LLAVA-1.5 [41] 7B	+0.4	-0.6	+0.7	+0.1	-8.4	+0.7	+3.8	-18.7	+1.6	+7.6	+3.6	+5.5	+3.9	+15.4	+3.7	+0.0	-3.7	+7.1
LLAVA-NEXT [36] (Mistral 7B)	+3.7	+1.1	+1.4	+3.4	-29.4	+10.0	+2.4	-5.5	+3.4	+3.5	-2.2	+2.0	-0.3	+8.8	+0.5	+0.0	-23.9	+4.5
LLAVA-NEXT [36] (Vicuna 7B)	+0.4	-3.7	+0.8	+2.9	-27.0	+10.2	+1.0	-11.1	+2.0	+5.0	-2.2	+3.1	+14.7	+3.9	+12.4	-0.0	-7.6	+5.4
LLAVA-OV [37] (Qwen2 0.5B)	+8.2	+8.7	+3.9	+7.5	-18.5	+16.9	+16.8	-17.1	+16.7	+25.6	+38.7	+20.6	+25.4	+10.9	+19.4	+0.1	-4.0	+19.6
LLAVA-OV [37] (Qwen2 7B)	+36.4	+22.8	+23.9	+2.4	-5.4	+7.7	+16.0	-7.1	+10.7	+10.1	+3.5	+14.8	-0.0	-4.3	+0.5	+0.0	-8.7	+22.3
PHI-3-VISION [1]	+6.0	-4.5	-1.8	+1.4	-8.4	+5.0	+19.7	-6.5	+12.2	+5.4	+2.1	+3.8	+10.4	+5.0	+6.4	-0.1	-0.5	+9.0
QWEN2VL [60] 2B	+7.5	+7.2	+5.5	+21.8	+0.7	+19.4	+10.2	-1.3	+11.1	+9.8	+6.9	+6.5	+38.7	+22.3	+24.5	+0.0	+17.8	+5.5
QWEN2VL [60] 7B	+7.9	+3.1	+4.4	+38.2	+16.1	+42.2	+13.6	-1.8	+11.8	+7.7	+3.5	+4.9	+3.3	+9.5	+4.3	-0.1	+13.7	+17.2

Table 14. Relative performance variation with dataset-specific prompts on six datasets. TI stands for text inclusion, LI for Llama inclusion, and CS for concept similarity.



(a) Default prompt (b) Ask for generic (c) Ask for specific (d) Dataset prompts

Figure 12. Types of model predictions when using the generic and specific prompts and the dataset-specific prompts. Blue indicates correct and specific and correct but generic predictions, red indicates wrong but specific and wrong and generic mistakes.

Model	Correct		Wrong	
	Specific	Generic	Specific	Generic
Be generic				
IDEFICS2 [34] 8B	-5.9	+11.3	-1.4	-4.0
INSTRUCTBLIP [19] Vicuna 7B	-9.9	+2.2	+0.5	+7.3
INTERNVL2 [12, 13] 2B	-5.3	+4.8	-1.2	+1.7
INTERNVL2 [12, 13] 4B	-9.3	+14.5	-3.9	-1.4
INTERNVL2 [12, 13] 8B	-20.6	+21.9	-5.5	+4.2
LLAVA-1.5 [41] 7B	-20.0	+3.8	-3.9	+20.0
LLAVA-NEXT [36] (Mistral 7B)	-8.0	+11.8	-3.2	-0.5
LLAVA-NEXT [36] (Vicuna 7B)	-17.8	+22.2	-4.2	-0.2
LLAVA-OV [37] (Qwen2 0.5B)	-7.1	-2.8	-0.8	+10.7
LLAVA-OV [37] (Qwen2 7B)	-2.9	+2.9	+0.6	-0.5
PHI-3-VISION [1]	-11.9	+6.2	+0.1	+5.6
QWEN2VL [60] 2B	-10.5	+0.6	+0.3	+9.6
QWEN2VL [60] 7B	-36.4	+19.5	+1.4	+15.5
Be specific				
IDEFICS2 [34] 8B	-3.2	-3.3	+0.7	+5.8
INSTRUCTBLIP [19] Vicuna 7B	+2.1	-6.7	+0.6	+4.0
INTERNVL2 [12, 13] 2B	-2.4	+0.2	+0.7	+1.5
INTERNVL2 [12, 13] 4B	-1.1	-5.4	+1.8	+4.7
INTERNVL2 [12, 13] 8B	-2.3	-3.5	+3.0	+2.8
LLAVA-1.5 [41] 7B	-2.9	-7.0	+0.8	+9.1
LLAVA-NEXT [36] (Mistral 7B)	-2.3	-3.4	+2.6	+3.1
LLAVA-NEXT [36] (Vicuna 7B)	-4.5	-1.7	+2.0	+4.2
LLAVA-OV [37] (Qwen2 0.5B)	-4.5	-1.8	-1.4	+7.7
LLAVA-OV [37] (Qwen2 7B)	+2.1	-5.6	+1.9	+1.6
PHI-3-VISION [1]	+0.4	-3.6	+2.3	+0.9
QWEN2VL [60] 2B	+2.7	-3.8	+3.7	-2.7
QWEN2VL [60] 7B	-14.4	+0.8	+1.4	+12.1
Dataset-specific				
IDEFICS2 [34] 8B	+14.9	-32.5	+7.0	+10.6
INSTRUCTBLIP [19] Vicuna 7B	+9.1	-10.7	+5.6	-4.0
INTERNVL2 [12, 13] 2B	+2.6	-3.9	+1.7	-0.4
INTERNVL2 [12, 13] 4B	+2.2	-5.5	+2.6	+0.7
INTERNVL2 [12, 13] 8B	-0.1	-3.6	+2.5	+1.2
LLAVA-1.5 [41] 7B	+5.6	-7.0	+3.6	-2.2
LLAVA-NEXT [36] (Mistral 7B)	+0.3	-7.1	+4.6	+2.2
LLAVA-NEXT [36] (Vicuna 7B)	+3.4	-8.6	+4.3	+1.0
LLAVA-OV [37] (Qwen2 0.5B)	+29.2	-8.9	+1.6	-21.9
LLAVA-OV [37] (Qwen2 7B)	+13.6	-14.8	+8.7	-7.5
PHI-3-VISION [1]	+5.5	-7.6	+2.6	-0.5
QWEN2VL [60] 2B	+13.6	-5.8	+1.1	-8.8
QWEN2VL [60] 7B	+15.0	-10.1	+2.1	-7.0

Table 15. Gains on the types of model prediction when instructing the models to be more generic/specific, and when using dataset-specific prompts techniques on six datasets, *i.e.*, DTD, FGVCAircraft, Flowers102, Food101, OxfordPets, StanfordCars.

Model	Caltech101			DTD			Flowers102			OxfordPets			UCF101		
	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS	TI	LI	CS
Zero-shot chain-of-thought															
INTERNVL2 [12, 13] 2B	+0.3	-0.5	+4.0	+2.3	+18.5	+5.3	-3.5	-7.8	+1.3	+4.9	-14.9	+5.3	+6.3	+5.3	+6.3
INTERNVL2 [12, 13] 4B	-5.1	+2.2	-0.7	+1.4	+29.3	+3.1	+0.9	+16.0	+2.7	+3.0	+6.9	+4.9	+4.6	+23.9	+4.6
INTERNVL2 [12, 13] 8B	-2.3	+2.2	+0.5	+4.0	+20.7	+3.5	-1.7	+13.8	+0.1	+2.3	-9.7	+4.2	+4.3	+14.9	+3.2
QWEN2VL [60] 2B	+1.8	+3.6	+5.5	+6.5	+25.4	+8.2	-1.9	+5.4	+2.0	+6.8	+3.6	+6.7	+5.7	+23.2	+7.9
QWEN2VL [60] 7B	-2.9	+5.0	+1.5	+6.5	+21.5	+5.5	-4.9	+13.6	-1.5	+7.1	+17.0	+7.7	+2.3	+23.9	+5.4
LlamaV-o1 multi-round prompt															
INTERNVL2 [12, 13] 2B	+0.4	+2.4	+4.1	+3.0	+7.8	+4.1	-2.6	-7.0	+1.9	+9.9	-18.8	+8.3	+5.5	+11.4	+5.0
INTERNVL2 [12, 13] 4B	-2.7	-4.3	-3.3	-0.5	+4.1	-0.7	+0.2	-9.1	-0.5	+3.4	-31.1	-5.2	+4.1	+3.9	+2.6
INTERNVL2 [12, 13] 8B	-1.5	-2.0	+1.3	+3.2	+8.9	+3.6	-2.5	+3.0	-2.2	+8.2	-15.5	+7.6	+5.4	+3.1	+2.8
QWEN2VL [60] 2B	+0.8	+0.3	+4.8	+6.3	+8.1	+6.7	-5.8	-7.7	-4.1	+27.6	+0.1	+19.5	+7.9	+19.1	+9.7
QWEN2VL [60] 7B	-1.0	-3.0	-1.5	+3.3	+2.4	+0.2	-6.8	-17.3	-10.3	+22.9	-3.4	+16.4	+3.3	+10.9	+4.2
LLaVA-COT prompt															
INTERNVL2 [12, 13] 2B	-0.6	+6.4	+4.1	+1.6	+19.8	+3.9	-3.8	+11.2	+1.2	+7.9	-8.6	+6.7	+5.8	+11.9	+5.2
INTERNVL2 [12, 13] 4B	+0.1	+5.0	+2.0	+0.8	+23.7	+2.9	-4.9	+30.8	-3.0	+6.7	+7.7	+7.2	+3.6	+16.4	+3.7
INTERNVL2 [12, 13] 8B	-1.7	+5.8	+1.3	+0.4	+25.3	+2.8	-8.9	+44.7	-6.8	-2.9	+28.9	+1.3	+3.2	+18.8	+0.8
QWEN2VL [60] 2B	+1.6	+0.4	+5.2	+4.4	+10.3	+6.4	-8.9	-7.9	-5.4	+5.6	-8.0	+6.1	+9.1	+18.5	+10.6
QWEN2VL [60] 7B	+0.3	+3.0	+3.4	+0.3	+12.5	+4.5	-10.2	+8.8	-7.2	+8.5	+16.2	+9.6	+6.6	+22.6	+7.1

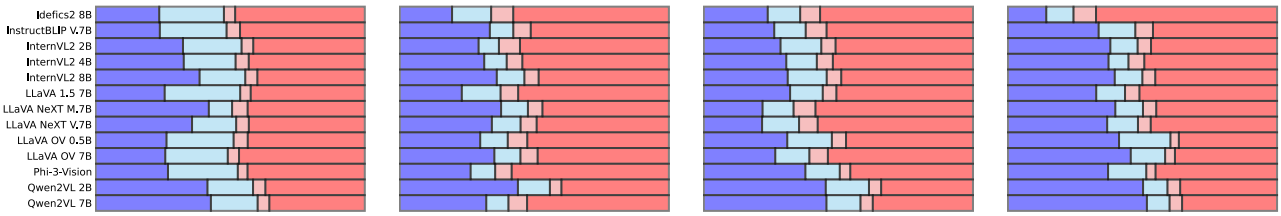
Table 16. Relative performance variation with chain-of-thought prompts on five datasets. TI stands for text inclusion, LI for Llama inclusion, and CS for concept similarity.

Model	Correct		Wrong	
	Specific	Generic	Specific	Generic
Zero-shot chain-of-thought				
INTERNVL2 [12, 13] 2B	+3.5	-5.5	+1.4	+0.6
INTERNVL2 [12, 13] 4B	+4.8	+9.7	-2.0	-12.5
INTERNVL2 [12, 13] 8B	+3.9	+2.1	-1.1	-4.9
QWEN2VL [60] 2B	+8.0	+3.1	-0.4	-10.8
QWEN2VL [60] 7B	+6.9	+8.6	-2.5	-13.1
LlamaV-o1 prompt				
INTERNVL2 [12, 13] 2B	+6.7	-8.8	+0.4	+1.7
INTERNVL2 [12, 13] 4B	+0.5	-9.8	+0.4	+9.0
INTERNVL2 [12, 13] 8B	+3.7	-6.2	+0.6	+1.9
QWEN2VL [60] 2B	+12.7	-8.6	+1.0	-5.1
QWEN2VL [60] 7B	+4.4	-6.3	+0.8	+1.0
LLaVA-CoT prompt				
INTERNVL2 [12, 13] 2B	7.3	-1.2	-0.9	-5.2
INTERNVL2 [12, 13] 4B	5.2	10.0	-2.3	-12.8
INTERNVL2 [12, 13] 8B	1.5	22.5	-2.8	-21.1
QWEN2VL [60] 2B	6.3	-4.2	0.3	-2.4
QWEN2VL [60] 7B	6.3	7.0	-2.0	-11.3
Reasoning models				
INTERNVL2.5 [11] 2B	-2.5	-7.6	+4.2	+6.0
INTERNVL2.5 [11] 4B	+4.7	-2.4	+4.4	-6.6
INTERNVL2.5 [11] 8B	+0.7	-0.3	+4.1	-4.5
QWEN2.5VL [4] 3B	+10.8	-6.5	+2.6	-6.9
QWEN2.5VL [4] 7B	+19.1	-9.4	+3.5	-13.2

Table 17. Gains on the types of model prediction when instructing the models to reason with chain-of-thought, and when using reasoning models on five datasets, *i.e.*, Caltech101, DTD, Flowers102, OxfordPets, UCF101.

Model	Correct		Wrong	
	Specific	Generic	Specific	Generic
List				
IDEFICS2 [34] 8B	-4.2	-9.5	+4.0	+9.7
INSTRUCTBLIP [19] Vicuna 7B	+9.6	-16.1	+1.5	+5.0
INTERNVL2 [12, 13] 2B	-3.1	-14.2	+3.5	+13.9
INTERNVL2 [12, 13] 4B	-1.4	-10.9	+0.9	+11.3
INTERNVL2 [12, 13] 8B	-2.6	-6.7	+0.8	+8.5
LLAVA-1.5 [41] 7B	-2.6	-13.7	+2.6	+13.6
LLAVA-NEXT [36] (Mistral 7B)	-4.4	+1.3	-0.4	+3.5
LLAVA-NEXT [36] (Vicuna 7B)	-1.6	-5.7	+1.3	+6.0
LLAVA-OV [37] (Qwen2 0.5B)	+3.6	-14.7	+2.0	+9.1
LLAVA-OV [37] (Qwen2 7B)	+9.3	-13.5	+2.2	+2.0
PHI-3-VISION [1]	-0.5	-16.8	+2.5	+14.9
QWEN2VL [60] 2B	+2.4	-5.0	-0.3	+3.0
QWEN2VL [60] 7B	-10.7	-9.1	+2.6	+17.2
Caption				
IDEFICS2 [34] 8B	+0.1	-12.0	+3.2	+8.6
INSTRUCTBLIP [19] Vicuna 7B	+2.3	-13.1	+2.8	+8.0
INTERNVL2 [12, 13] 2B	-4.0	-6.6	+1.0	+9.6
INTERNVL2 [12, 13] 4B	-2.1	-7.9	+1.1	+8.9
INTERNVL2 [12, 13] 8B	-7.4	-2.5	+0.9	+9.0
LLAVA-1.5 [41] 7B	+6.4	-15.9	+1.2	+8.4
LLAVA-NEXT [36] (Mistral 7B)	-20.2	+2.9	+2.4	+15.0
LLAVA-NEXT [36] (Vicuna 7B)	-14.2	-2.7	+2.2	+14.6
LLAVA-OV [37] (Qwen2 0.5B)	+4.7	-8.4	+0.0	+3.7
LLAVA-OV [37] (Qwen2 7B)	+0.7	-10.6	+2.7	+7.2
PHI-3-VISION [1]	+10.9	-13.2	+0.3	+2.0
QWEN2VL [60] 2B	+3.8	-0.9	-0.1	-2.8
QWEN2VL [60] 7B	+2.7	-4.7	+0.2	+1.8
Describe				
IDEFICS2 [34] 8B	-9.4	-13.9	4.2	19.1
INSTRUCTBLIP [19] Vicuna 7B	9.9	-11.2	1.6	-0.3
INTERNVL2 [12, 13] 2B	5.6	-11.6	0.6	5.4
INTERNVL2 [12, 13] 4B	4.6	-11.9	1.0	6.2
INTERNVL2 [12, 13] 8B	1.1	-6.7	0.3	5.3
LLAVA-1.5 [41] 7B	7.2	-17.4	1.5	8.7
LLAVA-NEXT [36] (Mistral 7B)	-2.1	1.6	-0.9	1.4
LLAVA-NEXT [36] (Vicuna 7B)	1.1	-5.0	0.0	3.9
LLAVA-OV [37] (Qwen2 0.5B)	15.0	-5.8	-2.1	-7.0
LLAVA-OV [37] (Qwen2 7B)	19.8	-10.5	-0.5	-8.8
PHI-3-VISION [1]	10.3	-11.8	-0.2	1.6
QWEN2VL [60] 2B	8.7	-7.9	0.1	-0.9
QWEN2VL [60] 7B	8.8	-8.9	0.3	-0.2

Table 18. Gains on the types of model prediction when instructing the models with multi-label prompts on five datasets, *i.e.*, Caltech101, DTD, Flowers102, OxfordPets, UCF101.



(a) Default prompt (b) List objects (c) Caption (d) Describe

Figure 13. Types of model predictions when using multi-label prompts. Blue indicates correct and specific and correct but generic predictions, red indicates wrong but specific and wrong and generic mistakes.

Model	Datasets										
	C101	DTD	ESAT	FGVC	FLWR	FOOD	PETS	CARS	S397	U101	Avg.
Text inclusion											
INTERNVL2.5 [10] 2B	55.8	12.6	**12.6**	1.5	10.9	17.0	8.7	0.0	16.3	13.7	14.9
INTERNVL2.5 [10] 4B	55.6	10.9	12.1	0.9	12.2	24.9	14.6	0.0	23.7	14.9	17.0
INTERNVL2.5 [10] 8B	56.4	12.1	8.4	3.0	16.8	29.7	7.2	**0.1**	24.7	13.8	17.2
QWEN2.5VL [4] 3B	62.1	13.9	1.6	18.8	49.7	44.2	38.9	0.0	30.7	18.0	27.8
QWEN2.5VL [4] 7B	**65.6**	**16.7**	4.4	**32.7**	**56.1**	**54.9**	**65.1**	0.0	**33.6**	**21.5**	**35.1**
Closed-world baselines											
CLIP [49]	87.1	52.6	42.7	27.2	76.9	89.9	88.1	76.2	65.6	72.7	67.9
SigLIP [68]	93.6	60.8	42.1	46.0	88.2	94.1	95.4	92.3	69.9	82.1	76.5
Llama inclusion											
INTERNVL2.5 [10] 2B	76.8	49.2	**47.2**	55.4	42.4	34.2	39.2	49.3	49.3	51.1	49.4
INTERNVL2.5 [10] 4B	77.1	48.7	42.6	61.4	43.3	52.0	49.4	49.8	63.1	53.6	54.1
INTERNVL2.5 [10] 8B	78.4	48.9	45.5	59.1	51.2	53.2	48.2	60.6	62.7	52.7	56.1
QWEN2.5VL [4] 3B	81.4	58.1	6.3	58.9	71.5	68.7	51.4	58.9	78.9	58.8	59.3
QWEN2.5VL [4] 7B	**84.5**	**59.8**	12.6	**69.6**	**75.2**	**76.4**	**71.0**	**71.2**	**81.1**	**67.0**	**66.8**
Closed-world baselines											
CLIP [49]	87.1	52.6	42.7	27.2	76.9	89.9	88.1	76.2	65.6	72.7	67.9
SigLIP [68]	93.6	60.8	42.1	46.0	88.2	94.1	95.4	92.3	69.9	82.1	76.5
Semantic similarity											
INTERNVL2.5 [10] 2B	49.5	25.2	31.4	21.4	26.7	33.5	22.4	41.8	39.8	41.5	33.3
INTERNVL2.5 [10] 4B	51.7	26.7	**31.7**	20.9	29.4	41.6	27.4	41.9	46.5	43.5	36.1
INTERNVL2.5 [10] 8B	**53.2**	27.1	29.5	21.4	32.1	42.2	24.2	42.9	**47.0**	43.2	36.3
QWEN2.5VL [4] 3B	51.8	27.4	12.3	28.9	45.4	48.0	31.4	50.9	**47.0**	43.2	38.6
QWEN2.5VL [4] 7B	48.8	**28.2**	18.9	**36.5**	**47.4**	**52.4**	**41.1**	**55.0**	**47.0**	**44.2**	**42.0**
Closed-world baselines											
CLIP [49]	90.8	69.9	67.7	66.7	83.4	93.7	91.8	80.5	92.2	83.3	82.0
SigLIP [68]	97.8	75.6	63.1	80.0	92.0	96.4	96.8	98.1	83.1	89.6	87.3
Concept similarity											
INTERNVL2.5 [10] 2B	78.0	46.2	**59.9**	33.1	47.8	53.5	39.7	50.0	59.5	61.4	52.9
INTERNVL2.5 [10] 4B	77.3	44.8	57.1	31.1	49.3	61.7	45.8	48.3	66.1	61.8	54.3
INTERNVL2.5 [10] 8B	77.7	45.4	52.7	31.5	54.2	64.5	41.9	49.2	66.3	62.2	54.6
QWEN2.5VL [4] 3B	81.8	51.3	23.8	52.4	72.6	73.2	62.1	64.7	70.8	62.9	61.6
QWEN2.5VL [4] 7B	**85.8**	**53.2**	41.3	**68.4**	**79.7**	**79.6**	**77.3**	**68.4**	**74.1**	**67.1**	**69.5**
Closed-world baselines											
CLIP [49]	90.8	69.9	67.7	66.7	83.4	93.7	91.8	80.5	92.2	83.3	82.0
SigLIP [68]	97.8	75.6	63.1	80.0	92.0	96.4	96.8	98.1	83.1	89.6	87.3

Table 19. OW results of reasoning models on ten datasets. Higher is better, **bold** indicates best. Note that the Llama inclusion for CLIP closed-world equals the textual inclusion scores.