Multimodal Large Language Models as Image Classifiers

Nikita Kisel Illia Volkov Klara Janouskova* Jiri Matas

Visual Recognition Group, Czech Technical University in Prague

kiselnik@fel.cvut.cz, volkoill@cvut.cz, janoukl1@fel.cvut.cz, matas@fel.cvut.cz

*Corresponding author

Abstract

Multimodal Large Language Model (MLLM) classification performance depends critically on evaluation protocol and ground truth quality. Studies comparing MLLMs with supervised and Vision-Language Models (VLMs) report conflicting conclusions, and we show these conflicts stem from protocols that either inflate or underestimate performance. Across the most common evaluation protocols, we identify and fix key issues: model outputs that fall outside the provided class list and are discarded, inflated results from weak multiple-choice distractors, and open-world setting that underperforms only due to poor output mapping. We additionally quantify the impact of commonly overlooked design choices — batch size, image ordering, and text encoder selection — showing they substantially affect accuracy.

Evaluating on ReGT, our multilabel reannotation of 625 ImageNet-1k classes, reveals that MLLMs benefit most from corrected labels (up to +10.8%), substantially narrowing the perceived gap with supervised models. Much of the reported MLLM underperformance on classification is thus an artifact of noisy ground truth and flawed evaluation protocol rather than genuine model deficiency. Models less reliant on supervised training signals prove most sensitive to annotation quality. Finally, we show that MLLMs can assist human annotators: in a controlled case study, annotators confirmed or integrated MLLM predictions in approximately 50% of difficult cases, demonstrating their potential for large-scale dataset curation. This work is part of the Aiming for Perfect ImageNet-1k project, see klarajanouskova.github.io/ImageNet.

1. Introduction

Multimodal Large Language Models (MLLMs) [1, 2, 6, 8, 9, 20, 26, 27], a recent extension of Large Language Models (LLMs) [4, 11, 28, 30] into the visual domain, are capable of performing complex vision-language reasoning that integrates linguistic understanding with visual perception. MLLMs performance has been extensively benchmarked [13, 17, 22, 23], primarily assessing high-level multimodal reasoning and instruction following through multiple-choice question answering. MLLMs exhibit remarkable generalization and are increasingly used by both researchers and the general public. Use-cases range from safety-critical ones such as medical diagnostics [5, 19, 37] and dangerous species classification [12], to helping with mundane office work [15, 42].

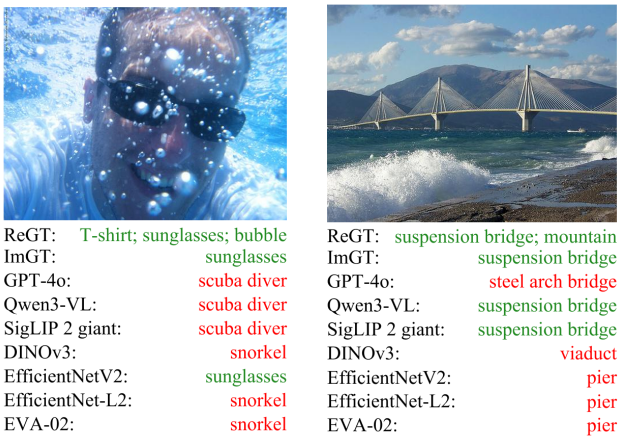

Figure 1. Challenging visual recognition cases from ImageNet-1k. ImGT: original single ImageNet label, ReGT: our reannotations. Predictions of representative models, including DINOv3, EfficientNetV2, EfficientNet-L2, EVA-02 (self-)supervised on ImageNet, are often wrong on such data. Correct predictions in green.

Classification performance provides a natural basis for comparing MLLMs with established computer vision methods, both supervised and VLMs (CLIP [29], SigLIP [35, 44]). Recent works have begun to explore this comparison [21, 40, 45]. However, depending on the evaluation protocol and task formulation, studies report conflicting conclusions: while some find that MLLMs perform substantially worse than classical vision models [45], others suggest performance comparable to VLMs [7, 21].

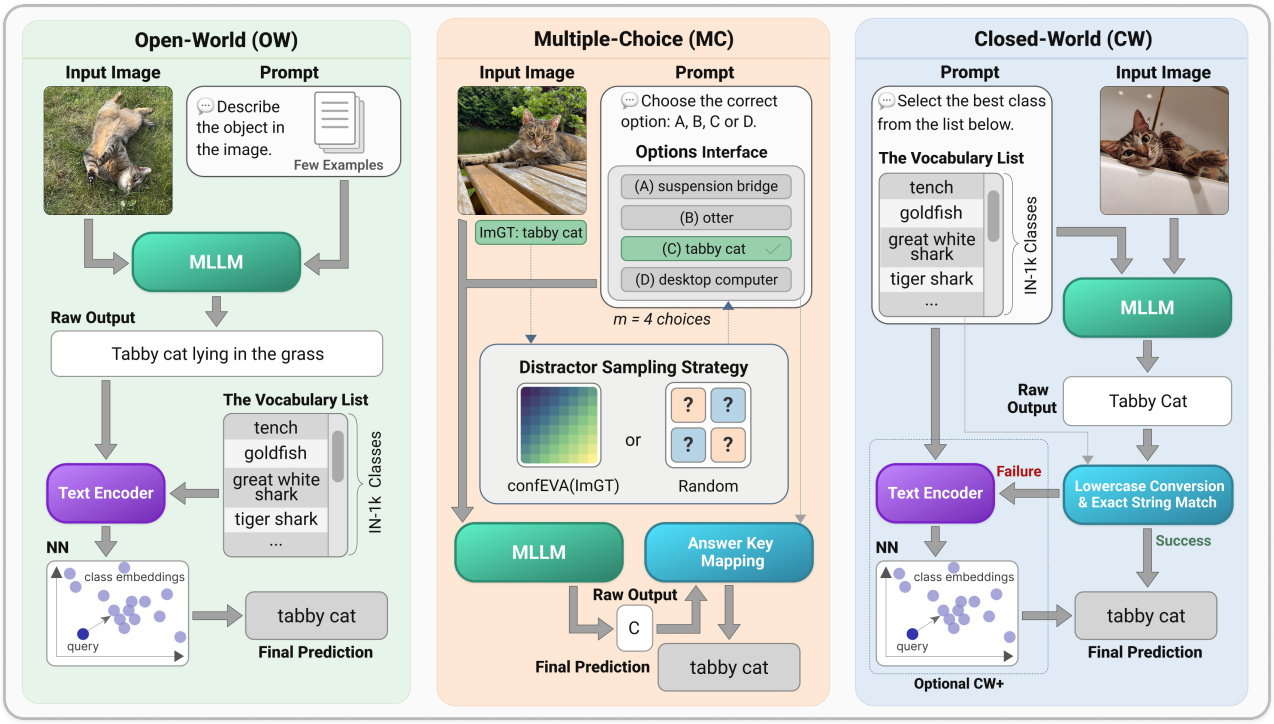

Figure 2. The three evaluated classification tasks — OW, MC, and CW(+) — described in Sec. 2.3

MLLMs produce free-form outputs, whereas image classification requires selecting a label from a predefined class set, so model outputs must be mapped to these classes. In the literature, this mapping has been achieved by formulating MLLM classification as one of the following tasks – Open-World (OW) [7, 45], Multiple-Choice (MC) question answering [21] and Closed-World (CW) [40, 45].

Open-World (OW) is closest to everyday MLLM use: the model generates a free-form image description, which is then mapped to dataset classes, either through simple heuristics (*i.e.*, text inclusion) [7, 45] or nearest-neighbor search in a text-embedding space [7]. Following Conti et al., who found embedding-based mapping to be the most effective, we adopt this strategy and further show that OW outperforms CW for half of the evaluated models. This contradicts prior work [45], which relied on string matching, indicating that previous OW limitations were due to the mapping method rather than the task itself.

Multiple-choice (MC) question answering is commonly used to benchmark MLLMs and is therefore often adopted for classification evaluation. The model selects from a set of candidate class names (up to 26 in prior work), including exactly one ground-truth label and several distractors. Distractors are sampled per-image, ranging from random to more challenging alternatives semantically close to the ground-truth label. Prior work reported that harder distractors cause only a "slight decrease" in performance [21]. We find this does not hold for newer models, where our confidence-interval-backed experiments show a larger drop under their sampling strategy. We also propose a distractor sampling method that further increases task difficulty. Yet, MC evaluation still inflates performance compared to other tasks, yielding an overly optimistic view of model ability.

The Closed-World (CW) task is designed to mimic classification with supervised models and VLMs. The model is prompted with candidate classes, ideally covering the full set of dataset classes. However, such a full-class formulation has been avoided due to input token limitations of older models [21, 45]; in our work, we are the first to use the full set of classes. As the number of candidate classes increases, out-of-prompt (OOP) predictions become more frequent, where the model generates a label that is not contained in the provided candidate list despite explicit instructions to select from it. While OOP predictions are theoretically possible in MC, they are not observed in practice due to the small number of candidates. In the literature, CW is typically avoided in favor of MC [21]; when it is used, OOP predictions are counted as incorrect [45]. We introduce CW+, which resolves OOP predictions by adopting the embedding-space nearest neighbour mapping from OW. With this approach, the main reason for avoiding CW is removed, making MC unnecessary unless the model is

constrained by input token length.

Even the best evaluation protocol is only as good as its ground truth. ImageNet-1k has long served as a foundation for pretraining and evaluating model performance. Thanks to its widespread adoption, its limitations are well-documented [3, 16, 25, 32, 36, 38, 43]. They go beyond simple image label errors (∼20% error rate in validation set [25]), including 15–21% of validation images containing multiple objects from different classes [3], making single-label evaluation unreliable, overlapping class definitions [38], distribution shifts between training and validation sets [16], and duplicate images [38]. Despite these limitations, ImageNet remains the de facto standard for evaluating visual recognition systems, including MLLMs [21, 40, 45].

We reannotate a substantial portion of the ImageNet-1k validation set (625 classes), mainly excluding challenging fine-grained animal categories that require extensive domain knowledge, greatly reducing annotation noise, ambiguities and other imperfections mentioned. The new labels, ReGT, are not yet public and thus unseen by any current model which enables controlled analysis of how supervised, self-supervised vision–language, and instruction-tuned MLLMs respond to improved ground truth.

The gap between top-performing MLLMs and supervised models is almost halved on the new labels. ReGT also partitions images into label categories based on agreement and multiplicity with ImGT—the original ImageNet-1k ground truth. Although supervised models remain strongest on average, self-supervised VLMs and MLLMs outperform them on images where ReGT disagrees with ImGT.

In our work we benchmark five MLLMs — the closed-source GPT-4o [27] and the open-source Qwen3-VL [41], LLaVA-OneVision [18], InternVL3.5 [6], and PaliGemma 2 [33] on the reannotated data. We also conduct preliminary experiments to study prompt design, reproducibility, encoder selection for CW+ and OW, and the influence of batch size and image order.

In a focused case study on the subset where ChatGPT disagreed with our ReGT, we conducted a second annotation pass in which annotators reviewed each image alongside anonymized predictions from GPT-4o, SigLIP 2, ImGT, and ReGT. For roughly half of these disputed cases with single-label, annotators either fully replaced or augmented the labels with GPT-4o's predictions, revealing substantial residual label noise even after an initial careful re-annotation. This demonstrates that MLLMs can serve as powerful annotation assistants.

To assess performance on fine-grained biological categories excluded from reannotation, we additionally evaluate on the expert-curated weasel-family subset from [16].

The main contributions are:

1. **Improved MLLM Image Classification Benchmark.** We systematically evaluate five MLLMs on ImageNet-1k across Closed-World, Open-World, and Multiple-Choice setups within a single framework, enabling direct comparison with vision-language and supervised baselines. We introduce CW+, a lightweight embedding-based post-processing that resolves out-of-prompt predictions without costly constrained decoding, enabling full 1,000-class Closed-World evaluation. To support fair assessment, we provide a new multilabel reannotation (ReGT) of 625 classes that holistically addresses known ImageNet labeling issues.
2. **Label Noise Sensitivity Across Learning Paradigms.** Leveraging ReGT, we show that MLLMs benefit most from corrected labels (up to +10.8%), substantially narrowing the perceived performance gap with supervised models. This reveals that much of the reported MLLM weakness on classification is an artifact of noisy ground truth rather than genuine model deficiency, and that models less reliant on supervised training signals are more sensitive to annotation quality.
3. **Evaluation Protocol Sensitivity.** We quantify the impact of commonly overlooked design choices—distractor selection, batch size, image ordering, output format, and text encoder—showing that they substantially affect reported accuracy (*e.g.* confusion-matrix distractors cause a 10–15% drop over random ones), calling into question prior results obtained under inflated conditions.
4. **MLLMs as Annotation Assistants.** In a controlled case study, human annotators confirmed or integrated the MLLM prediction in approximately 50% of difficult cases, demonstrating that MLLMs can effectively flag annotation errors and assist large-scale dataset curation.

2. Evaluation setup

2.1. Dataset and labels

All experiments are conducted on the ImageNet-1k dataset. Several prior works reannotated portions or even the entire validation set [3, 25, 36, 43]. However, each of these efforts addressed only a subset of the problems [16].

We introduce reannotated labels (ReGT) for 625 classes, holistically addressing known issues. We excluded fine-grained wildlife categories (except for dog breeds and species easily distinguishable without expert domain knowledge), as prior work [16, 24] showed them to be extremely challenging to annotate. To mitigate selection bias, we additionally evaluate on expert-provided *Mustelidae* reannotations [16] in the Supplementary.

Annotators, trained to recognize and avoid common issues, were asked to label all objects from ImageNet-1k classes (or indicate that none is present). To make it easier to remember all classes, the annotation tool displayed top-20 model predictions for each image. While potentially

Cat.	Definition	Cat.	Definition
A	$N \cup S \cup M$	S^+	$S \cap \{i : gt_i \in L_i\}$
N	$\{i : \lvert L_i\rvert = 0\}$	S^-	$S \cap \{i : gt_i \notin L_i\}$
S	$\{i : \lvert L_i\rvert = 1\}$	M^+	$M \cap \{i : gt_i \in L_i\}$
M	$\{i : \lvert L_i\rvert > 1\}$	M^-	$M \cap \{i : gt_i \notin L_i\}$

Table 1. Definitions of ImageNet-1k label categories based on ReGT and ImGT differences. For image i, gt_i is ImGT label, and L_i the set of ReGT labels. A, S, M correspond to all, single-label, and multi-label respectively, +/- denotes agreement/disagreement of ReGT and ImGT.

	A	S	S+	S−	M	M+	M−	N
#	31250	18071	16177	1894	11834	10756	1078	1345
%	100	57.8	51.8	6.1	37.9	34.4	3.5	4.3

Table 2. The number of images in each label category (#) and their fraction among reannotated images (%). Images with multiple labels (M) account for roughly 40 % of image in the selected 625 classes. Approximately 5 % of the images have no correct corresponding ImageNet-1k label (N).

introducing bias, this prevented many label omissions found in ImGT. The annotators collaborated via group chat to refine class definitions and challenging cases.

Label categories. The categorization of images based on the number of labels in ReGT and agreement with ImGT is provided in Tab. 1.

2.2. Evaluation metric

For evaluation on ReGT, we adapt the top-1 ReaL accuracy [3]. In ReaL, a prediction is counted as correct if it matches any of an image ground-truth labels, making it suitable for multilabel images. We treat predictions on images in N (those without any valid ground-truth label) as correct, since no meaningful label supervision exists for them. We also account for semantically equivalent classes[1] (*e.g.* "notebook computer" vs. "laptop"), a known issue in ImageNet [16, 38]. We treat each pair in a predefined equivalence set E as interchangeable, marking a prediction correct if it matches any label in the corresponding equivalence group. This applies to both ImGT and ReGT. Formally, for the image i we define a set of admissible labels:

$$L_i^* = L_i \cup \{b \mid \exists a \in L_i,\ \{a, b\} \in E\}. \tag{1}$$

Then, the accuracy over a set of images I is

$$\text{Acc} = \frac{1}{|I|} \sum_{i \in I} \alpha_i, \quad \alpha_i = \begin{cases} 1, & p_i \in L_i^* \text{ or } |L_i| = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where p_i is the prediction for image i.

[1]Full list is in the Supplementary.

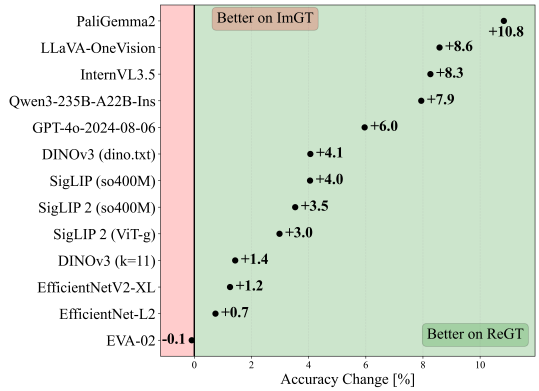

Figure 3. Classification accuracy on ImageNet-1k, change from the original (ImGT) to reannotated labels (ReGT).

2.3. Tasks

Task 1: Open-World (OW). The model is asked to classify the image without any predefined set of categories. A few prediction examples are given in the prompt to define the desired level of granularity. The predicted output is then embedded via a text encoder and matched with the nearest embedded class name to give the final prediction.[2]

Task 2: Multiple-Choice (MC). The prompt specifies n choices for each image and the model is asked to select one as the output. The choices always contain 1 correct answer and $n - 1$ distractor answers. This task can be considered a special case of Task 1 with only a subset of the classes used, but the prompt is also different. We adopt the standard [13, 17, 22, 23] $n = 4$ in all experiments. We explore multiple variants, differing in the distractor sampling strategy. The 4 options always include either ImGT, ReGT, or both. The distractors can be selected at random, or based on the confusion matrix of the supervised EVA-02 model, denoted as `confEVA(c)`, to favour classes that are likely to be confused with the class `c`. It is ensured that the ImGT label distractors do not include any ReGT image labels. The evaluation is performed multiple times, each time varying the choices order and the random distractors.

Task 3: Closed-World (CW). The prompt provides a list of all 1000 dataset classes and the model is asked to output a single one from the list. Since many objects may be present in the image, the model is asked to select only one class name that best describes the main object. Both the prediction and the class names are converted to lowercase; if they do not match, the prediction is considered incorrect.

2.4. Model and class names overview

Evaluated models. We evaluate five MLLMs: closed-source GPT-4o [27] and open-source Qwen3-VL [41], LLaVA-OneVision [18], InternVL3.5 [6], and PaliGemma 2

[2]Prompts for all Tasks are available in the Supplementary.

Model	Task	ImGT	ReGT							Im ∩ Re
		A_{31250}	A_{31250}	S_{18071}	$S+_{16177}$	$S-_{1894}$	M_{11834}	$M+_{10756}$	$M-_{1078}$	
PaliGemma2-mix-28B/448	OW	37.11 (13)	47.94 +10.8, (13)	39.74 (13)	41.49 (13)	24.76 (13)	54.55 (13)	56.24 (13)	37.66 (5)	35.78
LLaVA-OneVision-72B-Chat	OW	62.00 (12)	70.58 +8.6, (12)	67.11 (12)	70.67 (12)	36.69 (4)	72.52 (12)	75.87 (12)	39.05 (3)	59.39
InternVL3.5-38B	CW+	64.31 (11)	72.57 +8.3, (11)	67.69 (11)	71.05 (11)	39.02 (1)	76.91 (11)	80.43 (11)	41.74 (1)	61.53
Qwen3-VL-235B-A22B-Inst	OW	68.74 (10)	76.68 +7.9, (10)	73.61 (10)	77.79 (10)	37.91 (3)	78.71 (10)	82.46 (10)	41.28 (2)	65.87
GPT-4o-2024-08-06	CW+	76.40 (9)	82.36 +6.0, (9)	81.11 (9)	86.12 (9)	38.28 (2)	82.27 (9)	86.74 (9)	37.66 (5)	72.79
DINOv3 ViT-L/16 (dino.txt)	CW	80.96 (8)	85.02 +4.1, (8)	84.57 (8)	90.26 (8)	36.01 (6)	84.00 (7)	88.69 (8)	37.29 (6)	76.79
SigLIP so400M/14-384	CW	82.40 (7)	86.45 +4.0, (6)	85.59 (7)	91.35 (7)	36.38 (5)	86.23 (5)	90.96 (6)	38.96 (4)	78.27
SigLIP 2 so400M/16-384	CW	83.30 (6)	86.83 +3.5, (4)	86.06 (4)	91.99 (6)	35.32 (7)	86.52 (4)	91.47 (5)	37.20 (7)	78.93
SigLIP 2 ViT/g-opt-384	CW	84.14 (4)	87.12 +3.0, (3)	86.40 (3)	92.55 (4)	33.90 (8)	86.75 (3)	91.88 (3)	35.53 (8)	79.61
DINOv3 ViT-7B k=11	CW	84.18 (5)	85.61 +1.4, (7)	85.63 (6)	92.48 (5)	27.14 (11)	83.95 (8)	89.29 (7)	30.71 (11)	79.07
EfficientNetV2-XL	CW	85.28 (3)	86.53 +1.3, (5)	85.96 (5)	92.67 (3)	28.56 (9)	85.88 (6)	91.74 (4)	27.37 (12)	80.07
EfficientNet-L2	CW	88.17 (2)	88.91 +0.7, (2)	87.94 (2)	95.05 (2)	27.24 (10)	89.12 (2)	94.77 (2)	32.75 (9)	82.69
EVA-02 ViT-L-14-448	CW	90.13 (1)	90.04 -0.1, (1)	89.08 (1)	96.37 (1)	26.82 (12)	90.38 (1)	96.35 (1)	30.80 (10)	84.37

Table 3. Recognition accuracy on label categories of ImageNet-1k defined in Tab. 1. The highest-performing task (OW/CW+) is selected for the MLLMs. Increase or decrease in classification accuracy on ReGT w.r.t. ImGT. Model (rank)s are also shown. All models achieve higher accuracy on single-label images. The more dependence there is on the training data, the less the model benefits from the reannotation. The Im ∩ Re column reports the intersection of the two correctness sets: images on which the model is correct under ImGT and images on which it is correct under ReGT. This value indicates the overlap between these two correctness subsets, which are largely different.

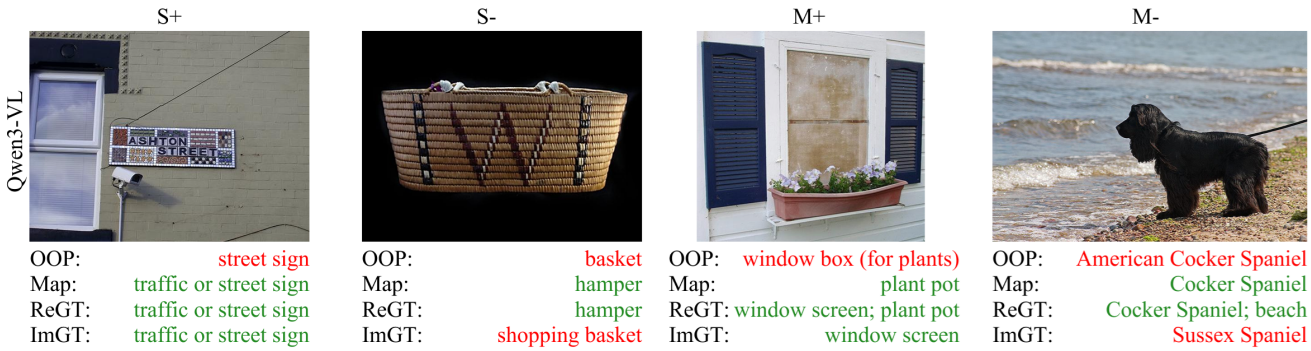

Figure 4. To address MLLMs out-of-prompt "OOP" predictions, often referred to as *hallucinations* [21, 45], we map model outputs that fall outside the provided class list to the nearest in-prompt class using the best model-specific encoder. Examples where the mapping is correct are shown. Columns correspond to the label categories from which the images were sampled. The row shows Qwen3-VL predictions, for which 38.75% OOP predictions are correctly mapped. Correct or incorrect image label.

[33]. GPT-4o was selected instead of the recent GPT-5 model due to ongoing API issues that make reliable evaluation difficult. Qwen3-VL was chosen as a representative leading open-source MLLM, while LLaVA-OneVision, InternVL3.5, and PaliGemma 2 cover a broader range of open-source architectures[3]. For the CW experiment, we additionally evaluate two families of VLMs: SigLIP so400M/14-384 [44] and SigLIP 2 in two sizes (so400M/16-384 and ViT/g-opt-384) [35]; DINOv3 as a zero-shot text-guided classifier 'dino.txt' (ViT-L/16) and as a k-NN classifier using ImageNet-1k training set (ViT-7B, k=11); and the supervised classifiers EfficientNet-L2 (trained to be robust to label noise), EfficientNetV2-XL [34] (trained on ImageNet21k), and transformer-based EVA-02 (best-performing model on ImageNet-1k [39]).

Text encoder. Text-embeddings (required for OW and CW+ are computed using the best model-specific text encoder in the main text. The standard mean over multiple prompt templates is used following [29]. Ablations for alternative encoders are provided in the Suppl., Tab. 17.

Class names. For CW and MC, we adopt the set of handcrafted ImageNet-1k class names introduced in Kisel et al. [16]. Although we find that they contain imperfections, these are the most up-to-date version available for the dataset. It is possible to learn "optimal" class representations [46, 47] and thereby increase accuracy, but this approach would no longer be zero-shot, which contradicts the goal of our MLLMs classification evaluation.

[3] See Tab. 7 in the Supplementary for details.

		ImGT	ReGT						
	Task	A_{31250}	A_{31250}	S_{18071}	$S+_{16177}$	$S-_{1894}$	M_{11834}	$M+_{10756}$	$M-_{1078}$
PaliGemma 2	OW	37.11	47.94	39.74	41.49	24.76	54.55	56.24	37.66
LLaVA-OV	OW	62.00 +18.6	70.58 +17.8	67.11 +17.5	70.67 +18.7	36.69 +7.1	72.52 +20.3	75.87 +22.0	39.05 +4.1
	CW+	52.66 +9.3	62.82 +10.1	59.54 +10.0	62.32 +10.4	35.80 +6.2	63.61 +11.4	65.84 +11.9	41.37 +6.4
	CW	43.40	52.75	49.59	51.94	29.57	52.20	53.92	34.97
InternVL3.5	OW	59.23 -4.9	68.18 -4.2	62.44 -5.1	65.78 -5.1	33.90 -5.1	73.33 -3.5	76.66 -3.7	40.07 -1.4
	CW+	64.31 +0.2	72.57 +0.2	67.69 +0.2	71.05 +0.2	39.02 +0.0	76.91 +0.1	80.43 +0.1	41.74 +0.3
	CW	64.14	72.42	67.51	70.84	39.02	76.78	80.32	41.47
Qwen3-VL	OW	68.74 +4.9	76.68 +5.3	73.61 +3.5	77.79 +3.7	37.91 +1.9	78.71 +8.7	82.46 +8.9	41.28 +7.0
	CW+	66.74 +2.9	74.43 +3.0	72.90 +2.8	76.95 +2.8	38.23 +2.2	73.86 +3.8	77.53 +3.9	37.29 +3.0
	CW	63.86	71.39	70.15	74.14	36.06	70.04	73.61	34.32
GPT-4o	OW	70.92 -4.5	77.58 -3.7	76.59 -3.7	81.08 -4.3	38.23 +1.2	76.55 -4.1	80.75 -4.4	34.69 -1.6
	CW+	76.40 +1.0	82.36 +1.1	81.11 +0.8	86.12 +0.7	38.28 +1.2	82.27 +1.6	86.74 +1.6	37.66 +1.4
	CW	75.40	81.29	80.32	85.39	37.06	80.65	85.10	36.27

Table 4. Recognition accuracy for CW(+) and OW tasks. PaliGemma 2 results are reported only for the OW setup, since it is not feasible to perform CW with all class names included in a single prompt. Changes relative to the Closed World (CW) baseline are indicated as increase or decrease. The CW+ setup extends the CW task by applying the same text-embedding-space output mapping used in the Open World (OW) task, addressing OOP predictions in CW that would otherwise be considered incorrect. Results are obtained using the best-performing model-specific encoder; for a comparison of all encoders, see Tab. 17. As expected, accuracy in CW+ is higher than in standard CW. Notably, LLaVA-OV and Qwen3-VL outperform their CW results in the OW setup, an unexpected trend that contrasts prior findings on MLLMs [21, 45]. In contrast, InternVL3.5 and GPT-4o achieve the highest performance in CW+.

Model	A	S	S+	S−	M	M+	M−	N
InternVL3.5	0.9	0.8	0.8	0.8	0.8	0.7	1.2	2.5
GPT-4o	5.3	4.0	3.4	8.7	6.0	5.6	9.9	16.4
Qwen3-VL	10.5	9.3	8.7	13.8	10.9	10.6	14.8	24.2
LLaVA-OV	26.8	24.1	23.6	28.1	29.0	28.8	31.2	42.5

Table 5. MLLMs out-of-prompt (OOP) rate, the (%) of cases when the output is none of the classes specified, per label categories. Models hallucinate more on 1. multilabel images than single label ones, 2. images that do not contain their ImGT in ReGT (S- and M-), 3. images from N category (contain no ImageNet labels in ReGT). On the last, the hallucination rate is highest for all models.

3. Results

We compare MLLMs, VLMs, and vision-only architectures on ImageNet-1k with ImGT and ReGT. The results reveal differences in robustness and label dependence, with ReGT reducing the performance gap and an embedding-based mapping further mitigating hallucination errors.

Preliminary experiments assessing the impact of batching (size, ordering and composition) and prompting with either class names or class IDs are provided in the Supplementary.

3.1. Closed-World

Results on ImGT are reported in Tab. 3, first column. Among MLLMs, PaliGemma 2 scores 37 %, while LLaVA-OV (62 %), InternVL3.5 (64 %) and Qwen3-VL (69 %) score closely. GPT-4o dominates at 76 %. Despite this impressive result, there is still a gap between GPT-4o and vision-specific self-supervised models, with the leading SigLIP 2 reaching 84 %. Supervised models still outperform all others, with EfficientNetV2-XL and EfficientNet-L2 scoring 85 % and 88 %, respectively, while EVA-02 achieves the highest accuracy of 90 %.

Results on ReGT yield the largest gains for MLLMs: PaliGemma 2 improves by 10.8 %, LLaVA-OV by 8.6 %, InternVL3.5 by 8.3 %, Qwen3-VL by 7.9 %, and GPT-4o by 6.0 %. The performance of VLMs still shows a solid improvement of 3–4.1 %. Supervised models improve negligibly: EfficientNetV2-XL gains 1.3 %, EfficientNet-L2 improves by 0.7 % and EVA-02 decreases by 0.1 %. This confirms the intuition that the more a model relies on the training data, the smaller the reannotation benefit. The results also show VLMs and MLLMs do not overfit the official ImageNet-1k validation labels, as reflected on the S- and M- subsets, where the top ranks go to self-supervised models rather than supervised ones. Deltas between ImGT and ReGT are presented in Fig. 3.

Notably, the gap between MLLMs and supervised models got reduced significantly, for the better performing GPT-4o by roughly 6 %. As shown in Tab. 2, our non-OOD reannotated labels do not contain the ground-truth label in 9.6 % of images (S−, M−), which is close to the error rate

		ImGT	ReGT					
	Distractors	A_{625}	A_{625}	S_{352}	$S{+}_{316}$	$S{-}_{36}$	M_{240}	$M{+}_{221}$	$M{-}_{19}$
GPT-4o	ImGT + random	99.62 ±0.07	90.95 ±0.05	89.49 ±0.07	99.67 ±0.07	0.09 ±0.18	91.85 ±0.10	99.75 ±0.11	0.00 ±0.00
	ImGT + `confEVA(ImGT)`	90.66 ±0.22	85.12 ±0.21	84.27 ±0.26	93.87 ±0.29	0.00 ±0.00	84.31 ±0.39	91.56 ±0.42	0.00 ±0.00
	ReGT + `confEVA(ReGT)`	51.35 ±0.21	84.26 ±0.37	92.45 ±0.36	93.90 ±0.31	79.66 ±1.50	70.09 ±0.89	70.78 ±0.83	62.14 ±3.36
	ImGT + ReGT + random	91.77 ±0.26	94.89 ±0.13	94.79 ±0.16	99.67 ±0.07	51.97 ±1.35	94.31 ±0.27	99.80 ±0.09	30.56 ±2.97
	ImGT + ReGT + `confEVA(ImGT,ReGT)`	88.31 ±0.29	92.52 ±0.15	91.59 ±0.20	96.36 ±0.15	49.73 ±1.27	92.86 ±0.31	98.16 ±0.20	31.24 ±3.30
	ImGT + 999 (Closed World (CW))	74.69 ±0.19	81.32 ±0.18	79.87 ±0.19	84.49 ±0.21	39.25 ±0.82	80.89 ±0.35	84.72 ±0.36	36.33 ±0.58

Table 6. 4-way multiple-choice (MC) task results on a randomly sampled subset of 625 images (one per class) with 95 % confidence intervals (CI). The function `confEVA()` generates challenging distractors based on the confusion matrix of EVA-02. The Closed World result corresponds to 999 distractors. Only GPT-4o results are shown here, as the trend of gradually declining accuracy with more challenging distractors is similar across models. Full multiple-choice task results for all evaluated MLLMs are provided in Tab. 18.

Figure 5. Images from the second annotation pass, where annotators 1. preferred GPT-4o over ReGT (left), 2. kept part of ReGT without adding GPT-4o (right), and 3. combined ReGT with GPT-4o predictions (middle). "ReReGT" indicates second pass reannotations. Annotators in the second pass preserved correct ReGT or added GPT-4o predictions while also correcting erroneous first-pass labels.

of the best-performing supervised model EVA-02.

Out-Of-Prompt. MLLMs still often predict texts outside of the provided prompt class names despite clear instructions not to, an effect known as hallucinations [21, 45], which we investigate by label category in Tab. 5.

NN in text embedding space mapping (CW+). To address this, we encode model predictions into the text embedding space (similarly to the OW setup) and assign each to its nearest class name embedding (CW+, Tab. 4). The performance of CW+ consistently outperforms pure CW across all subsets and models, with the largest gains on the more challenging S- and M- subsets. The accuracy improvement achieved by LLaVA-OV (the highest among all evaluated models) is ≈54× greater than that of InternVL3.5 (the model with the lowest improvement) when evaluated against ImGT, and ≈67× greater when evaluated using ReGT. The difference is explained by LLaVA-OV's consistently higher OOP rate (Tab. 5), which also corresponds to higher correct mapping rates[4]. Examples of successful mappings are shown in Fig. 4.

3.2. Multiple-Choice

Results in Tab. 6 reveal the commonly adopted setup with random distractors significantly inflates performance, while harder distractors reduce it by 10–15 %. When both ImGT and ReGT are provided, models select between them at similar rates on challenging single label images ($S-$), where confidence intervals are also wider, as they are for $M-$.

[4]See Tab. 14 in the Supplementary for a comprehensive breakdown.

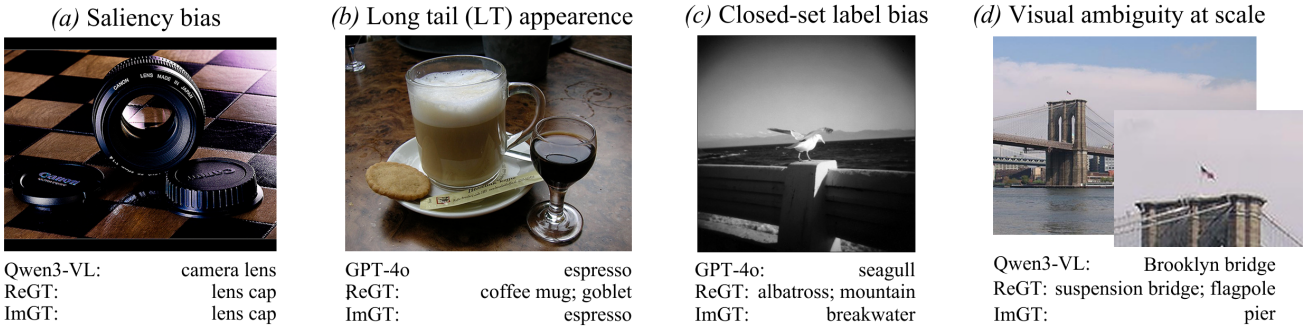

Figure 6. Challenging classification cases. *(a)* Dominant objects do not always match the target class; e.g., a "lens" $\notin$ ImageNet-1k labels. *(b)* Appearance variations hinder recognition of long-tail content, e.g., espresso served in a goblet vs. a cup. *(c)* Fixed label sets force assignment even for out-of-distribution objects (e.g., "seagull" $\notin$ ImageNet-1k). *(d)* Recognizability depends on scale and context; small or ambiguous objects are harder to identify.

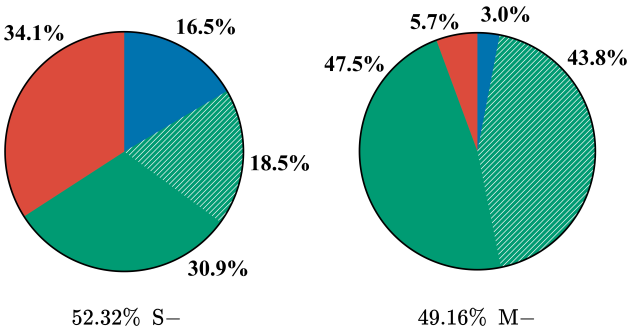

Figure 7. Second annotation pass results on images where GPT-4o gave a valid prediction and disagreed with ReGT for S− and M− label categories (52.32% and 49.16% of images in these categories, respectively; challenging since ImGT $\notin$ ReGT). Color coding: Annotators ■ preferred the GPT-4o prediction (ReGT was incorrect); ■ preserved correct ReGT without adding GPT-4o prediction; ■ combined GPT-4o prediction with ReGT (GPT-4o $\notin$ ReGT before); ■ made other changes (GPT-4o & ReGT $\notin$ ReReGT).

3.3. Open-World

Tab. 4 indicates that overall performance is noticeably lower than in the CW setup for InternVL3.5 and GPT-4o , whereas it is higher for LLaVA-OV and Qwen3-VL . The latter contrasts with previous studies [21, 45], which report that accuracy in the CW setup generally exceeds that of the OW for various MLLMs - a trend we also observe for InternVL3.5 and GPT-4o. Consistent with the CW results, all models show sizable improvement on ReGT compared to ImGT.

3.4. Case study: ChatGPT vs. humans

We conducted a second annotation pass on the challenging S− and M− label categories[5]. It revealed that GPT-4o predictions were frequently adopted either fully or in combination with ReGT (see Fig. 7). At the same time, a non-negligible fraction of cases highlighted inconsistencies in human corrections, suggesting that GPT-4o can serve as a useful assistive tool for error detection and verification.

Annotation challenges. Fig. 6 illustrates common failure cases in image classification and dataset annotation.

[5]See Sec. B in the Supplementary for full experimental setup.

4. Conclusions

We presented an evaluation of both closed- and open-source MLLMs on ImageNet-1k, enabled by a new large-scale re-annotation (ReGT) that reduces label noise and resolves long-standing ambiguities. The results showed MLLMs perceived performance is the most affected by incorrect ground truth. GT corrections narrowed the apparent performance gap w.r.t. supervised and self-supervised vision models, while also revealing strong MLLMs sensitivity to task formulation and distractor selection. The ReGT labels further expose how model families differ in their robustness to mislabeled and multilabel images, and demonstrate that MLLMs can meaningfully assist human annotators in a controlled curation pipeline. Although closed-source systems retain an advantage under strict Closed-World prompting, this gap largely disappears in Open-World settings.

We showed both the promise and the pitfalls of current MLLMs for visual recognition, underscoring the need for high-quality benchmarks and careful integration of model assistance in dataset construction.

Acknowledgments. Nikita Kisel and Illia Volkov were supported by EquiLibre Technologies, Klara Janouskova by the CTU student grant SGS. The work was also supported by the Technology Agency of the Czech Republic within the SIGMA Programme project No. TQ28000003 and by the EC Digital Europe Programme project CEDMO 2.0 no. 101158609. The infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".

References

[1] Anthropic. Claude 3, 2025. [Large language model]. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1

[3] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet?, 2020. 3, 4

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared Kaplan et. al. Language models are few-shot learners, 2020. 1

[5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024. 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 1, 3, 4

[7] Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers, 2025. 1, 2, 5, 8

[8] Google DeepMind. Gemini, 2025. [Large language model]. 1

[9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, and Jae Sung Park et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 1

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[12] Yuze Du, Yingjia Wang, and Eric Zhao. Leveraging multimodal llms for plant species identification and educational insights, 2024. 1

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, and Jinrui Yang et al. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. 1, 4

[14] Manu Gaur, Darshan Singh S, and Makarand Tapaswi. Detect, describe, discriminate: Moving beyond vqa for mllm evaluation, 2024. 1

[15] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models, 2024. 1

[16] Nikita Kisel, Illia Volkov, Kateřina Hanzelková, Klara Janouskova, and Jiri Matas. Flaws of imagenet, computer vision's favourite dataset. In *The Fourth Blogpost Track at ICLR 2025*, 2025. 3, 4, 5, 2

[17] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 1, 4

[18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 3, 4

[19] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. 1

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[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, 2024. 1, 2, 3, 5, 6, 7, 8

[22] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, and Ji Zhang et al. Mibench: Evaluating multimodal large language models over multiple images, 2024. 1, 4

[23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, and Wangbo Zhao et al. Mmbench: Is your multi-modal model an all-around player?, 2024. 1, 4

[24] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity, 2022. 3

[25] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. 3

[26] OpenAI. Chatgpt (gpt-5), 2025. [Large language model]. 1

[27] OpenAI et al. Gpt-4o system card, 2024. 1, 3, 4

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 5, 8

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. 1

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 1, 5

[32] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020. 3

[33] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang

Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer, 2024. 3, 5

[34] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021. 5

[35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1, 5

[36] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks, 2020. 3

[37] Tao Tu, Shekoofeh Azizi, Danny Driess, and Mike Schaekermann et. al. Towards generalist biomedical ai, 2023. 1

[38] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet, 2022. 3, 4

[39] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 5

[40] Wenhao Wu, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang, and Jingdong Wang. Gpt4vis: What can gpt-4 do for zero-shot visual recognition?, 2024. 1, 2, 3

[41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and Chujie Zheng et al. Qwen3 technical report, 2025. 3, 4, 5

[42] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 1

[43] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels, 2021. 3

[44] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 1, 5

[45] 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?, 2024. 1, 2, 3, 5, 6, 7, 8

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022. 5