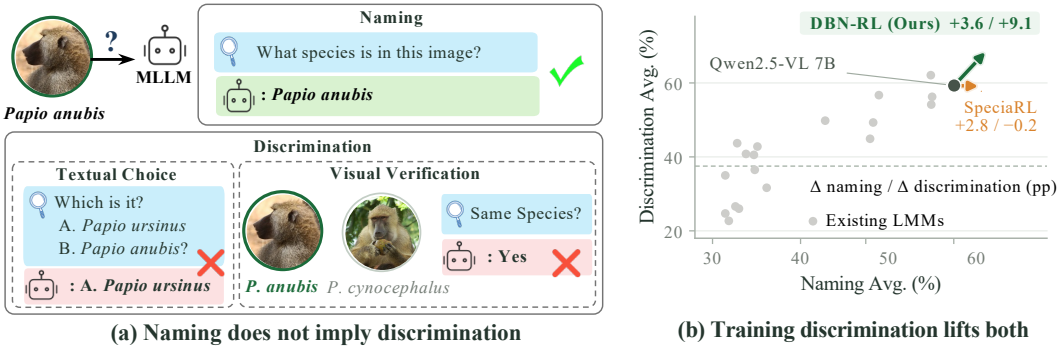
BEYOND NAMING: RE-BENCHMARKING OPEN-WORLD CLASSIFICATION WITH DISCRIMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world classification seeks to recognize objects without a predefined candidate set. With MLLMs, this ability is realized and evaluated through free-form naming, based primarily on whether the generated name is correct and sufficiently specific. However, a model may name a category correctly while failing to distinguish it from visually similar classes. We introduce **DOWC (Discrimination-Aware Open-World Classification)**, a benchmark that jointly evaluates naming and discrimination through textual and visual discrimination tasks. Our evaluation reveals that base models exhibit substantially weaker discrimination than naming, particularly in purely visual settings. To address this gap, we propose **DBN (Discriminate Before You Name)**, which introduces an explicit local class comparison before naming. DBN first constructs a comparison neighborhood from self-proposed plausible classes, selects the candidate best supported by the image, and uses this selection to guide naming. We instantiate DBN in two forms: **DBN-RL** internalizes this process through supervised warm-up and reinforcement learning, with rule-based rewards coupling answer correctness and selection consistency, whereas **DBN-TF** applies it training-free to frozen models through coarse-to-fine candidate proposal and confidence-based selection. Across five base models, DBN-RL and DBN-TF improve discrimination by **+11.0** and **+9.4** points and naming by **+7.8** and **+7.1** points, respectively. These results show that explicit local discrimination provides a stronger basis for open-world recognition.

1 INTRODUCTION



(a) Naming does not imply discrimination **(b) Training discrimination lifts both**

Figure 1: **Discrimination is an overlooked dimension of open-world classification with MLLMs.** (a) QWEN2.5-VL 7B names *Papio anubis* correctly but fails to distinguish it from *Papio ursinus*, while OWC evaluation checks only the name. (b) Training discrimination improves naming too.

Conventional image classification (Krizhevsky et al., 2012; He et al., 2016; Dosovitskiy et al., 2020), which has underpinned much of modern computer vision's progress, assumes a fixed and known label space: given an image, the model selects one class from a predefined set. Multimodal Large language models (MLLMs) (Liu et al., 2023; Chen et al., 2024c; Wang et al., 2024) relax this assumption by naming the object in an image directly in natural language, without a candidate list, enabling open-ended recognition. Open-world classification (OWC) (Conti et al., 2025) introduced a benchmark for evaluating this open-world naming capability, measuring whether the generated

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

name is correct. SpeciaRL (Angheben et al., 2026) further treated fine-grained naming as a trainable capability, using reinforcement learning to encourage more specific predictions. Together, these efforts organize both evaluation and training around the emitted name—whether it is correct and sufficiently specific. Accordingly, current OWC evaluations primarily assess naming.

Naming alone, however, does not fully characterize a model's recognition ability. Indeed, research in cognitive science shows that the perceived similarity structure among objects remains largely consistent across speakers of different languages despite substantial differences in naming, suggesting that object recognition and naming in humans are not perfectly coupled (Malt et al., 1999). This imperfect coupling calls into question treating the emitted name as the sole criterion for open-world classification. A classification decision must both identify the target class and distinguish it from plausible alternatives. In closed-world classification, this requirement is built into the prediction rule: an argmax over the fixed label space is correct only when the target class outranks every alternative, including confusable neighbors. Cross-entropy reinforces this ordering during training by increasing the target class's score relative to those of competing classes. Consequently, the same decision both accepts the target class and rejects its alternatives, making discrimination implicit in both training and evaluation. In generative open-world classification, however, a correct name does not guarantee reliable discrimination among confusable alternatives. Unlike the class-level argmax in closed-world classification, autoregressive generation produces a name token by token and exposes no explicit comparison among confusable classes. Fig. 1 (a) makes this gap concrete: on the same image, the model correctly names the baboon as *Papio anubis*, yet selects *Papio ursinus* when asked to choose between the two species. Current naming-only evaluation nevertheless scores the model as correct on this image despite its failure to reject the confusable neighbor. Discrimination is therefore an overlooked dimension of open-world classification: correct naming does not guarantee it, and current OWC benchmarks do not evaluate it.

To measure this missing dimension, we introduce **Discrimination-Aware Open-World Classification (DOWC)**, a benchmark extending OWC with an explicit discrimination axis. Given an image x and its ground-truth class c, DOWC evaluates both *naming*—whether the model outputs c or a semantically equivalent name—and *discrimination*—whether it identifies c over a confusable class c^*. Discrimination is evaluated by varying how the confusable class is presented, as a textual name or a visual instance, and how the decision is elicited, through verification or choice, yielding four tasks across fine-grained datasets. We further construct DOWC-Taxo, which varies the taxonomic rank of the competing classes, from family to genus to species, to provide a controlled difficulty axis for studying how discrimination changes as the class distinction becomes finer. DOWC enables us to examine whether correct naming is accompanied by reliable discrimination on the same images. Our measurements reveal two facts. First, on average, base models achieve reasonable naming accuracy but exhibit weak discrimination, especially when the confusable class is presented as a visual instance. Second, discrimination is strongly task-dependent: on the same images, choosing between given textual candidates is substantially more accurate than naming freely (Sec. 4).

These findings suggest explicit candidate comparison can make fine-grained discrimination more accessible. A small set of plausible candidates turns an otherwise implicit decision into a local class-level comparison, providing a focused basis for recognition. Building on this insight, we propose **Discriminate Before You Name (DBN)**, a framework that strengthens discrimination by incorporating this local comparison into the generation process. To realize this local comparison, DBN constructs the comparison neighborhood from plausible class hypotheses generated by the model itself, reflecting how the model interprets the image. The model then selects one candidate from this neighborhood, making the class-level discrimination explicit. The resulting local discrimination then guides the model's open-world recognition. We instantiate this framework in two ways. **DBN-RL** internalizes the process through supervised warm-up followed by reinforcement learning, using self-generated training data and rule-based rewards enforce answer correctness and consistency with the selected candidate. **DBN-TF** realizes the same process at inference time with a frozen model, using coarse-to-fine candidate proposal and confidence-based selection; it is therefore training-free and plug-and-play. We evaluate DBN on DOWC using five base models. Although naming accuracy varies across these models, their discrimination remains limited, especially when confusable classes are presented as visual instances. DBN consistently improves discrimination, yielding average gains of 11.0 percentage points for DBN-RL and 9.4 points for DBN-TF over the corresponding base models. DBN also improves open-world naming, with average gains of 7.8 and 7.1 points,

respectively. These results suggest that explicitly comparing plausible, confusable candidates helps resolve fine-grained class ambiguity and provides a more reliable basis for open-world recognition.

Our main contributions are summarized as follows.

- We identify discrimination as an overlooked dimension of open-world classification with MLLMs, and empirically show that correct naming does not imply reliable discrimination.
- We introduce DOWC, a benchmark that evaluates naming and discrimination together. It defines four discrimination tasks over five fine-grained datasets, and its variant DOWC-Taxo controls discrimination difficulty through taxonomic distance.
- We propose DBN (Discriminate Before You Name), which turns the comparison implicit in token-by-token generation into an explicit local comparison among self-proposed candidates. The framework admits both a training-based and a training-free realization.
- We show that strengthening discrimination also improves naming. Across five base models, DBN raises discrimination by 11.0 points and naming by 7.8 points on average; training that targets naming alone leaves discrimination unchanged.

2 DOWC: Discrimination-Aware Open-World Classification

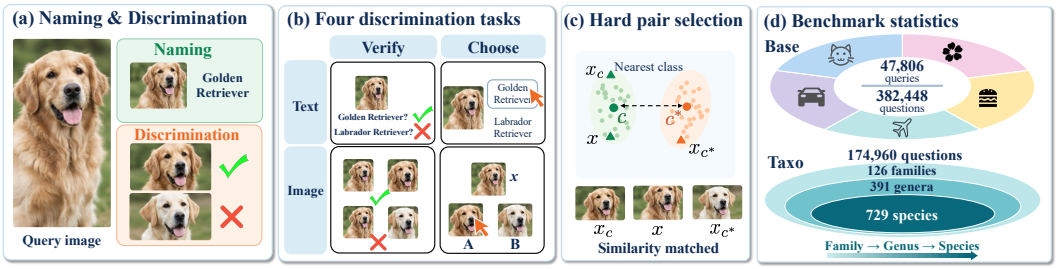


Figure 2: **Overview of the DOWC benchmark.** (a) DOWC evaluates open-world naming and discrimination. (b) Four discrimination tasks combine text/image modalities with verification/choice protocols. (c) Hard pairs are constructed by selecting the nearest confusable class and matching positive and negative instances in similarity to the query. (d) Statistics for DOWC and DOWC-Taxo, which evaluates discrimination at the family, genus, and species levels.

2.1 Formulation and Task Design

Naming and discrimination. Given an image x with ground-truth class c, we decompose open-world recognition into two independently testable conditions. *Naming* requires the model to output c or a semantically equivalent name. *Discrimination* requires the model to accept c and reject a confusable class c^*. Naming follows the standard OWC protocol, whereas DOWC evaluates discrimination in both textual and visual modalities.

Task design. As illustrated in Fig. 2 (b), we organize discrimination evaluation along two dimensions: the modality in which the confusable class is presented and the protocol used to test the model. ***First***, the confusable class can be presented textually or visually. In the textual modality, the model judges whether class names c and c^* match the query image x. In the visual modality, it judges whether the query image matches instance images x_c and x_{c^*}, without using class names. ***Second***, acceptance and rejection can be tested separately or jointly. The *verification* protocol evaluates them with separate queries, $\langle x, c\rangle$ and $\langle x, c^*\rangle$. The *choice* protocol presents both candidates jointly as $\langle x, c, c^*\rangle$ and asks the model to choose the correct one. Verification directly reflects the definition of discrimination, whereas choice is robust to constant-answer strategies.

Combining the two dimensions yields four tasks: Text-Verify (TV), Image-Verify (IV), Text-Choose (TC), and Image-Choose (IC). For each query image x, all four tasks use the same ground-truth and confusable classes (c, c^*); the image-based tasks additionally use the same instance pair (x_c, x_{c^*}). Their construction is described in Sec. 2.2.

2.2 Confusable-Pair Construction and Metrics

For each query image x with ground-truth class c, DOWC constructs a confusable class c^* and, for image-based tasks, a pair of instances (x_c, x_{c^*}). As illustrated in Fig. 2(c), the construction selects the hardest confusable class and matches the similarity of the positive and negative instances.

Selecting the confusable class c^*. We define confusability in visual feature space, since the discrimination tasks are grounded in image content rather than class-name similarity. For each evaluation dataset, we extract image features using a frozen DINOv3 ViT-B/16 encoder Siméoni et al. (2025) and represent each class by the mean of its image features. The confusable class is the class whose prototype is most similar to that of the ground-truth class:

$$c^* = \arg\max_{c' \neq c} \mathrm{sim}(\mu_c, \mu_{c'}), \tag{1}$$

where μ_c denotes the prototype of class c. All four discrimination tasks for a query image use the same pair (c, c^*). The main benchmark uses the nearest class, corresponding to the hardest setting; DOWC-Taxo varies the difficulty through taxonomic distance (Sec. 2.3).

Selecting the visual instances. The image-based tasks additionally require one instance from each class. We select the negative instance as the example from c^* that is most similar to the query:

$$x_{c^*} = \arg\max_{x_i \in c^*} \mathrm{sim}(x, x_i), \qquad s_{\mathrm{neg}} = \mathrm{sim}(x, x_{c^*}). \tag{2}$$

This gives the hardest rejection example. We then select the positive instance from c whose similarity to x is closest to s_{neg}:

$$x_c = \arg\min_{x_i \in c,\ x_i \neq x} \left| \mathrm{sim}(x, x_i) - s_{\mathrm{neg}} \right|. \tag{3}$$

Matching the positive and negative similarities removes a trivial similarity cue, so the resulting comparison more directly measures class-level discrimination. The image-based tasks use the same pair (x_c, x_{c^*}) for each query.

Metrics. For naming, we follow OWC Conti et al. (2025) and report Textual Inclusion (TI), Llama Inclusion (LI), Semantic Similarity (SS), and Concept Similarity (CS). For discrimination, Text-Verify (TV) and Image-Verify (IV) use paired accuracy: a query is counted as correct only when both the ground-truth class c is accepted and the confusable class c^* is rejected. Text-Choose (TC) and Image-Choose (IC) use answer accuracy over the two candidate orders. Prompt templates and auxiliary metrics are provided in App. B.

2.3 Datasets and Benchmark Statistics

DOWC-Base. DOWC-Base follows the data splits of OWC and retains five fine-grained datasets: Oxford Pets, Flowers102, Food101, FGVC-Aircraft, and Stanford Cars. These datasets contain 536 classes in total. We exclude the prototypical and non-prototypical groups because their classes are generally well separated and rarely provide meaningful confusable alternatives. The 47,806 query images produce 47,806 naming questions and 382,448 discrimination questions, with eight discrimination questions generated per query image.

DOWC-Taxo. DOWC-Taxo introduces controllable discrimination difficulty through biological taxonomy. Unlike DOWC-Base, where each class is paired with its nearest confusable class, DOWC-Taxo constructs discrimination questions at three taxonomic ranks: species, genus, and family. Query images are reused across three ranks, allowing model performance to be compared under different levels of class granularity. We select 729 species from iNat21 (Van Horn et al., 2021), spanning 391 genera and 126 families, with 10 images per species. Naming is evaluated at the species level, yielding 7,290 naming questions. At each taxonomic rank, the 7,290 query images produce 58,320 discrimination questions; across the three ranks, DOWC-Taxo contains 174,960 discrimination questions. At each rank, confusable classes and visual instance pairs are constructed following the procedure in Sec. 2.2. Statistics and settings are provided in App. B.

3 DBN: Discriminate Before You Name

Overview. DBN makes local discrimination explicit before answering by comparing self-generated confusable candidates in a local neighborhood around the query (Fig. 3). We instantiate this frame-

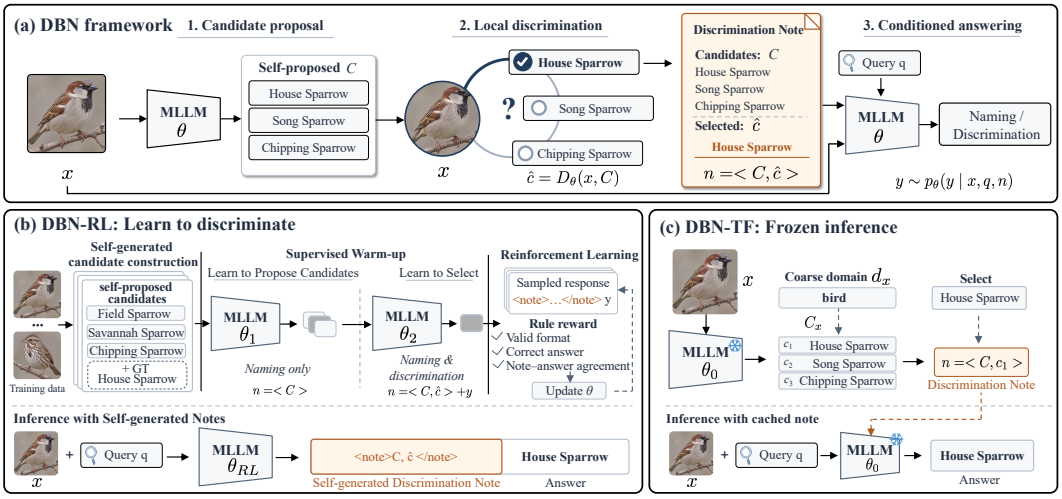


Figure 3: **Overview of Discriminate Before You Name (DBN).** (a) DBN constructs a local neighborhood of confusable candidates, performs local discrimination, and conditions answering on the resulting discrimination note. (b) DBN-RL learns the procedure through self-generated candidate construction, supervised warm-up, and rule-based reinforcement learning. (c) DBN-TF realizes the same procedure with a frozen model through coarse-to-fine candidate proposal, confidence-based selection, and cached note-assisted answering.

work as DBN-RL, which learns the procedure during training, and DBN-TF, which executes it with a frozen model at inference time.

3.1 DBN Framework

From implicit to explicit discrimination. Open-world recognition requires both naming an object and distinguishing it from its confusable neighbors. An MLLM may produce the correct name without reliably making this distinction. In open-world classification, the model generates an answer autoregressively, so comparisons among plausible classes remain implicit: each token is conditioned on the generated prefix and normalized over the full vocabulary, rather than being organized as an explicit comparison within a local neighborhood of confusable alternatives. Discrimination is therefore both a capability to evaluate and a mechanism that supports fine-grained naming. DBN makes this comparison explicit before producing the answer.

Framework factorization. DBN makes this comparison explicit by first constructing a local neighborhood of confusable hypotheses around the query image. Let x denote an image, q a textual query, and y the generated answer. A standard MLLM directly produces $y \sim p_\theta(y \mid x, q)$. DBN decomposes this process into three stages:

(i) *Candidate proposal:* the model generates K fine-grained candidate names that form a local candidate neighborhood, $C = \{c_1, \ldots, c_K\}, |C| = K$;

(ii) *Local discrimination:* the model selects one candidate from this neighborhood, $\hat{c} = \mathcal{D}_\theta(x, C), \hat{c} \in C$, where $\mathcal{D}_\theta$ denotes the model's local discrimination operator;

(iii) *Conditioned answering:* the model generates the answer conditioned on the query and the discrimination note,$n = \langle C, \hat{c} \rangle, y \sim p_\theta(y \mid x, q, n)$.

Here, n denotes the *discrimination note*, whose core content is the local candidate neighborhood C and the selected candidate $\hat{c}$.

Discussion. DBN is a model-agnostic framework rather than a particular architecture or prompt. It can be realized during training, where the model learns to construct candidates, select among them, and answer conditioned on the resulting note (DBN-RL; Sec. 3.2), or at inference time, where a frozen model performs the same steps through prompting without parameter updates (DBN-TF; Sec. 3.3).

3.2 DBN-RL: TRAINING-BASED INSTANTIATION

Let θ_0 denote the base MLLM before training. DBN-RL trains the model to emit a discrimination note n before its answer y within a single autoregressive generation, thereby modeling the joint distribution $p_\theta(n, y \mid x, q)$. Candidate content is generated from the model's own outputs, while ground-truth labels provide the supervision for the selected candidate and the final answer. No external model is used as a teacher or reward judge. Training consists of a supervised warm-up followed by reinforcement learning: the warm-up teaches the model to construct candidate neighborhoods and select among them, while reinforcement learning optimizes the self-generated note and answer with rule-based rewards.

Self-generated candidate construction. A training sample is a triple (x, q, t), where x is an image, q is a naming or discrimination query, and t is the target sequence consisting of a discrimination note followed by the correct answer to q. Let c_x^{gt} denote the ground-truth class of x. For each training image, the base model θ_0 is queried under multiple prompts to generate K fine-grained candidate names: $\mathcal{N}_{\theta_0}(x) = \{c_1, \ldots, c_K\}, |\mathcal{N}_{\theta_0}(x)| = K$. The candidates include the base model's free-form naming of x and its naming of the class visually closest to x. We construct the candidate set by adding the ground-truth class: $C_x = \{c_x^{\mathrm{gt}}\} \cup \mathcal{N}_{\theta_0}(x)$. The training note is $n_x = \langle C_x, c_x^{\mathrm{gt}} \rangle$, where the last entry indicates the selected candidate. We randomize the order of the candidates in C_x to prevent the model from relying on a fixed positional pattern. Because the candidate names are generated by θ_0, the resulting sets provide a model-specific approximation to its confusion structure. Each training image contributes samples for both naming and discrimination queries.

Supervised warm-up. The base model does not initially produce a candidate note before answering or select a candidate from it. We therefore use two stages of supervised fine-tuning. First, the model learns to emit the candidate part of the note. Let $n_x^C = \langle C_x \rangle$ denote this part. Using naming queries, we optimize

$$\mathcal{L}_1(\theta) = -\mathbb{E}_x \left[\log p_\theta \left(n_x^C \mid x, q_{\mathrm{name}} \right) \right]. \tag{4}$$

The resulting model, denoted by θ_1, learns to produce candidates before answering but has not yet been trained to select among them.

In the second stage, we construct training notes from the candidates generated by θ_1. Let $\widehat{C}_{\theta_1}(x)$ denote the candidate set generated by θ_1 for x. We add the ground-truth class only if it is absent:

$$\widetilde{C}_x = \widehat{C}_{\theta_1}(x) \cup \{c_x^{\mathrm{gt}}\}. \tag{5}$$

This matches the candidate distribution used for training with the distribution the model encounters when generating candidates at inference time. For naming and discrimination queries, we optimize

$$\mathcal{L}_2(\theta) = -\mathbb{E}_{(x,q)} \left[\log p_\theta \left(\widetilde{n}_x, y_q^* \mid x, q \right) \right], \qquad \widetilde{n}_x = \langle \widetilde{C}_x, c_x^{\mathrm{gt}} \rangle, \tag{6}$$

where y_q^* is the correct answer to query q. The resulting model is denoted by θ_2.

Reinforcement learning with rule-based rewards. The supervised warm-up teaches the model to imitate targets in which the selected candidate is always the ground-truth class. It establishes the output format but does not directly optimize the selections generated by the model itself. We therefore apply reinforcement learning to θ_2, operating on the notes and answers generated by the model. The reward encourages both a correct final answer and consistency between the answer and the selected candidate. Because answer validity and correctness can be verified by rules, this is a natural setting for reinforcement learning with verifiable rewards (RLVR) (Guo et al., 2025). We adopt GRPO (Shao et al., 2024), which samples a group of responses for each input and updates the policy using group-relative rewards without a separate value model. RLVR has also been applied successfully to vision tasks (Liu et al., 2025; Li et al., 2025b). During training, each input is a raw image-query pair without an additional prompt template, and both naming and discrimination queries are sampled.

For a response $o = (n, y)$, let n denote the generated discrimination note, y the final answer, and $\widehat{c}(n)$ the candidate selected in the note. Let y_q^* denote the correct answer to query q, and let $v(o) \in \{0, 1\}$ indicate that the note is well-formed and the answer is legally formatted. We define the rule-based reward as

$$r(o) = v(o) \Big(\lambda_a + \lambda_b \mathbb{1}[y = y_q^*] + (1 - \lambda_a - \lambda_b) \mathbb{1}[y \simeq \widehat{c}(n)] \Big), \tag{7}$$

where $y \simeq \widehat{c}(n)$ indicates agreement between the final answer and the selected candidate. For naming queries, y and $\widehat{c}(n)$ must denote the same class; for verification queries, the accept/reject

answer must be consistent with whether the queried class is selected; and for choice queries, the chosen option must correspond to the selected class.

Here λ_a gives a small credit for a valid response, λ_b weights answer correctness, and the remaining weight rewards consistency with the discrimination note. We use nonnegative weights with $\lambda_a + \lambda_b \leq 1$ and require $\lambda_b > 1 - \lambda_a - \lambda_b$, so that answer correctness is rewarded more strongly than consistency with an incorrect selection. During reinforcement learning, the note is not matched against a reference note; its content receives no direct supervision beyond validity and consistency with the final answer. The consistency term therefore discourages the model from bypassing the intermediate discrimination step. All reward terms are rule-based, and training does not rely on an external model as a judge.

3.3 DBN-TF: Training-free Instantiation

DBN-TF realizes the DBN factorization at inference time without additional training or parameter updates. Given an image, a frozen base model first scopes the candidate space with a coarse domain label, proposes a local neighborhood of fine-grained candidates, and selects the most confident candidate from this neighborhood. The candidate neighborhood and the selection are recorded as a discrimination note, which is then reused to answer naming and discrimination queries.

Constructing local discrimination. The frozen base model θ_0 proposes candidates with fixed prompts and selects among them by model confidence. As in DBN-RL, the candidates are generated by the model itself and form a local neighborhood of plausible, confusable hypotheses. We construct this neighborhood in three steps.

(1) Coarse-level scoping. We first ask the base model to identify the image's coarse category, denoted by d_x. This category scopes the subsequent proposal to related fine-grained classes.

(2) Fine-grained candidate proposal. Given the image and d_x, the model is prompted to propose K fine-grained candidate names, denoted by $C_x = (c_1, \ldots, c_K)$ and ordered by model confidence. These candidates form the local neighborhood for discrimination.

(3) Confidence-based selection. The highest-confidence candidate is selected as $\hat{c}_x$.

The resulting discrimination note, $n_x = \langle C_x, \hat{c}_x \rangle$, is generated once per image and reused across all queries involving that image. Prompt details are provided in App. D.

Note-assisted answering. For an image-conditioned query q, the frozen model generates $y \sim p_{\theta_0}(y \mid x, n_x \oplus q)$, where $\oplus$ denotes adding the discrimination note n_x to the query. The same note is generated once per image and reused across all associated naming and discrimination queries. For multi-image inputs, each image is provided with its corresponding note.

4 Experiments

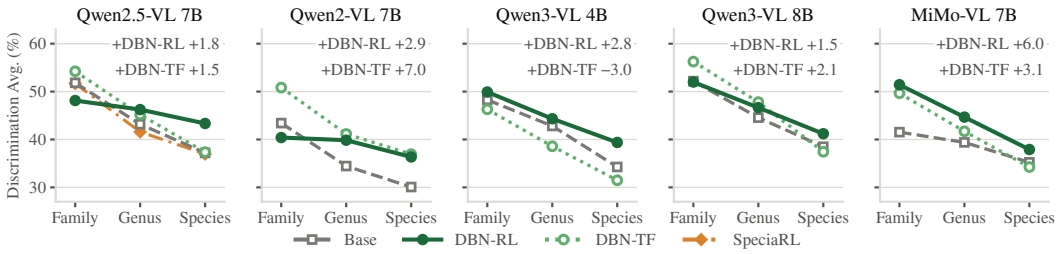


Figure 4: Discrimination on DOWC-Taxo by taxonomic rank (family, genus, species).

Models and baselines. We evaluate the open-source MLLMs from OWC (Conti et al., 2025) and SpeciaRL (Angheben et al., 2026), namely Idefics2 (Laurençon et al., 2024), InternVL2 (Chen et al., 2024c;b), InternVL2.5 (Chen et al., 2024a), LLaVA-OneVision (Li et al., 2024), Phi-3-Vision (Abdin et al., 2024), Qwen2-VL (Wang et al., 2024), and Qwen2.5-VL (Bai et al., 2025b), with CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) as retrieval references, together with three recent families: InternVL3.5 (Wang et al., 2025b), Qwen3-VL (Bai et al., 2025a), and MiMo-VL (Yue et al., 2025). We use reported naming results and measure all discrimination metrics ourselves. Both DBN variants are evaluated on Qwen2-VL 7B, Qwen2.5-VL 7B, Qwen3-VL 4B/8B, and MiMo-VL 7B; DBN-TF is additionally tested on Qwen3-VL 30B-A3B.

Table 1: Naming and discrimination results on DOWC (in %). Avg. denotes the mean of the four metrics in each block. Rows below the dashed line apply DBN to the base model above; subscripts indicate the gain over that base.

	Fine-grained										Very fine-grained									
	Naming					*Discrimination*					*Naming*					*Discrimination*				
Model	TI	LI	SS	CS	**Avg.**	TV	TC	IV	IC	**Avg.**	TI	LI	SS	CS	**Avg.**	TV	TC	IV	IC	**Avg.**
Large multimodal models																				
IDEFICS2 8B	3.0	49.9	38.0	41.7	33.1	49.8	76.1	14.8	50.1	47.7	0.0	67.0	29.6	33.6	32.5	35.4	67.5	6.1	49.6	39.7
INTERNVL2 2B	14.9	47.0	31.6	50.7	36.1	35.2	68.9	4.2	0.0	27.1	0.7	32.9	33.1	43.9	27.6	15.5	57.7	0.0	0.0	18.3
INTERNVL2 4B	16.2	44.4	32.0	52.0	36.1	42.8	73.1	5.5	0.0	30.3	1.7	36.8	33.8	44.2	29.1	25.1	63.5	3.1	0.0	22.9
INTERNVL2 8B	22.3	46.7	34.8	56.7	40.1	49.6	76.9	18.4	43.1	47.0	2.3	32.5	36.0	49.4	30.1	27.8	64.9	21.2	40.7	38.6
INTERNVL2.5 2B	12.2	38.6	27.5	47.0	31.3	38.9	69.9	4.2	0.0	28.2	0.8	52.4	31.6	41.5	31.6	20.3	60.4	3.9	0.0	21.2
INTERNVL2.5 4B	17.2	48.2	32.8	52.3	37.6	39.4	79.2	12.7	50.0	45.3	0.5	55.6	31.4	39.7	31.8	24.8	65.3	4.7	48.5	35.8
INTERNVL2.5 8B	17.9	50.9	32.8	53.5	38.8	44.6	80.1	10.9	4.5	35.0	1.6	59.9	32.1	40.4	33.5	30.9	69.0	13.1	0.0	28.3
INTERNVL3.5 8B	20.6	43.3	32.8	56.7	38.3	24.4	75.6	17.6	43.7	40.3	2.9	41.0	33.4	48.0	31.3	19.7	67.4	11.6	32.0	32.7
LLAVA-OV QWEN2 0.5B	6.0	42.7	38.5	43.3	32.6	40.8	70.9	0.5	21.8	33.5	0.6	65.6	30.5	37.1	33.4	19.2	52.7	0.6	1.6	18.5
LLAVA-OV QWEN2 7B	6.4	40.4	39.0	43.8	32.4	39.6	83.5	1.4	46.8	42.8	0.0	76.7	31.9	32.4	35.2	28.8	77.0	1.3	48.2	38.8
PHI-3-VISION	13.4	49.1	31.8	47.2	35.4	33.2	70.7	5.3	50.0	39.8	0.2	45.0	28.9	36.0	27.5	5.8	63.5	1.4	50.0	30.2
QWEN2-VL 2B	35.7	62.5	40.7	63.4	50.6	71.3	90.7	11.1	28.8	50.5	12.9	60.7	45.1	62.3	45.2	49.2	78.4	14.6	14.9	39.3
QWEN2.5-VL 3B	44.3	63.9	41.6	69.3	54.8	64.2	89.8	12.4	45.4	52.9	9.4	58.9	39.9	58.5	41.7	40.5	77.1	15.9	49.4	45.7
SpeciaRL (Qwen2.5-VL 7B)	62.7	74.4	49.2	81.1	66.8	65.8	92.6	35.7	63.3	64.3	24.9	63.8	50.5	75.4	53.6	47.1	80.1	26.8	61.7	53.9
MIMO-VL 7B	49.4	67.2	42.4	73.4	58.1	65.1	90.5	33.1	56.8	61.4	8.5	57.2	37.2	55.9	39.7	48.1	76.7	24.7	58.3	52.0
+ DBN-TF	53.7	69.6	52.7	75.5	62.9$_{+4.8}$	68.9	87.3	44.1	79.6	70.0$_{+8.6}$	17.3	67.5	50.2	69.5	51.1$_{+11.4}$	43.4	76.3	42.9	74.2	59.2$_{+7.2}$
+ DBN-RL	58.1	65.4	72.6	74.7	67.7$_{+9.6}$	75.6	91.6	57.4	80.4	76.3$_{+14.9}$	13.8	62.9	64.5	68.5	52.4$_{+12.7}$	45.9	78.8	45.2	71.9	60.4$_{+8.4}$
QWEN2-VL 7B	34.6	64.0	39.2	62.9	50.2	61.1	91.8	3.1	52.4	52.1	0.8	63.0	34.5	43.4	35.4	45.6	81.9	9.8	52.9	47.5
+ DBN-TF	53.0	69.9	55.6	74.7	63.3$_{+13.1}$	75.0	91.7	31.0	78.5	69.1$_{+17.0}$	26.3	78.0	56.8	73.0	58.5$_{+23.1}$	**69.3**	**86.4**	37.2	**78.2**	**67.8**$_{+20.3}$
+ DBN-RL	59.5	67.7	**75.5**	76.7	**69.9**$_{+19.7}$	68.5	89.9	**58.9**	81.9	74.8$_{+22.7}$	21.9	73.6	**70.3**	74.8	**60.2**$_{+24.8}$	50.0	84.8	**52.8**	74.4	65.5$_{+18.0}$
QWEN2.5-VL 7B	58.7	74.2	47.0	78.9	64.7	65.2	92.5	36.0	64.7	64.6	16.4	70.4	45.8	68.4	50.2	46.7	79.8	27.6	61.8	54.0
+ DBN-TF	64.2	76.2	49.9	**82.5**	68.2$_{+3.5}$	74.1	89.1	50.5	80.9	73.7$_{+9.1}$	**26.6**	67.8	51.2	**76.0**	55.4$_{+5.2}$	50.6	81.2	47.1	75.2	63.5$_{+9.5}$
+ DBN-RL	57.5	66.2	72.1	74.0	67.4$_{+2.7}$	70.1	92.3	55.5	77.2	73.8$_{+9.2}$	21.2	60.1	66.5	70.6	54.6$_{+4.4}$	55.0	81.6	42.8	72.1	62.9$_{+8.9}$
QWEN3-VL 4B	60.6	75.9	46.4	75.2	64.5	67.6	90.9	41.1	60.5	65.0	13.5	57.1	46.0	64.4	45.2	42.8	78.0	18.4	33.9	43.3
+ DBN-TF	**64.8**	73.9	51.1	80.6	67.6$_{+3.1}$	76.3	87.1	53.2	73.0	72.4$_{+7.4}$	14.2	64.6	46.8	67.3	48.2$_{+3.0}$	42.6	65.9	36.7	53.6	49.7$_{+6.4}$
+ DBN-RL	58.3	65.9	72.1	74.9	67.8$_{+3.3}$	71.6	90.0	56.8	83.0	75.3$_{+10.3}$	10.8	41.1	58.6	62.5	43.2$_{-2.0}$	30.9	74.0	44.5	73.4	55.7$_{+12.4}$
QWEN3-VL 8B	61.5	74.3	47.3	75.9	64.8	70.6	92.4	48.4	70.9	70.6	15.2	49.9	47.3	67.0	44.8	46.8	77.8	29.2	60.5	53.6
+ DBN-TF	63.4	70.7	50.7	79.6	66.1$_{+1.3}$	75.5	91.9	57.1	84.0	77.2$_{+6.6}$	16.0	59.8	47.1	67.9	47.7$_{+2.9}$	43.4	73.5	35.6	69.0	55.4$_{+1.8}$
+ DBN-RL	55.4	66.2	71.0	72.8	66.3$_{+1.5}$	72.7	91.1	51.9	83.0	74.7$_{+4.1}$	13.9	39.0	63.4	68.2	46.1$_{+1.3}$	39.4	72.4	36.7	70.7	54.8$_{+1.2}$
QWEN3-VL 30B-A3B	53.5	76.6	44.0	73.2	61.8	74.4	**94.2**	45.7	49.1	65.9	15.0	66.0	45.1	66.4	48.1	50.4	79.7	26.8	29.4	46.6
+ DBN-TF	49.1	**79.5**	42.8	71.7	60.8$_{-1.0}$	**82.2**	92.8	58.3	**84.1**	**79.4**$_{+13.5}$	17.0	**78.6**	44.1	64.4	51.0$_{+2.9}$	55.5	79.2	44.6	73.2	63.1$_{+16.5}$

Table 2: Naming performance on DOWC-Taxo.

Model	TI	LI	SS	CS	**Avg.**
QWEN2.5-VL 7B	5.8	24.0	56.2	56.5	35.6
+ DBN-TF	16.2	20.4	58.8	59.4	38.7$_{+3.1}$
+ DBN-RL	16.6	22.0	60.4	61.1	**40.0**$_{+4.4}$
+ SpeciaRL	16.7	18.4	53.4	54.8	35.8$_{+0.2}$
QWEN2-VL 7B	1.3	23.7	54.3	54.3	33.4
+ DBN-TF	8.9	20.1	57.9	58.2	36.3$_{+2.9}$
+ DBN-RL	11.4	20.5	58.1	58.5	37.1$_{+3.7}$
QWEN3-VL 4B	8.1	19.1	53.9	55.3	34.1
+ DBN-TF	11.5	16.5	55.2	56.1	34.8$_{+0.7}$
+ DBN-RL	11.9	15.6	57.0	57.8	35.6$_{+1.5}$
QWEN3-VL 8B	13.9	19.9	54.7	56.1	36.2
+ DBN-TF	17.7	19.3	57.1	58.0	38.0$_{+1.8}$
+ DBN-RL	15.3	18.6	57.5	58.4	37.4$_{+1.2}$
MIMO-VL 7B	12.2	13.1	47.4	48.8	30.4
+ DBN-TF	12.9	15.8	56.0	56.6	35.3$_{+4.9}$
+ DBN-RL	12.5	17.4	55.9	56.5	35.6$_{+5.2}$

Table 3: Ablation of DBN-RL training stages.

Model	TV	TC	IV	IC	**Avg.**
QWEN2.5-VL 7B	46.7	79.8	27.6	61.8	54.0
+ SFT warm-up	54.4	79.7	38.2	70.2	60.7$_{+6.7}$
+ RL (DBN-RL)	55.0	81.6	42.8	72.1	**62.9**$_{+8.9}$
QWEN2-VL 7B	45.6	81.9	9.8	52.9	47.5
+ SFT warm-up	49.6	82.7	50.4	70.2	63.2$_{+15.7}$
+ RL (DBN-RL)	50.0	84.8	52.8	74.4	**65.5**$_{+18.0}$
QWEN3-VL 4B	42.8	78.0	18.4	33.9	43.3
+ SFT warm-up	29.1	65.4[§]	35.4	68.8	49.7$_{+6.4}$
+ RL (DBN-RL)	30.9	74.0	44.5	73.4	**55.7**$_{+12.4}$
QWEN3-VL 8B	46.8	77.8	29.2	60.5	53.6
+ SFT warm-up	39.1	63.5[§]	34.0	69.1	51.4$_{-2.2}$
+ RL (DBN-RL)	39.4	72.4	36.7	70.7	**54.8**$_{+1.2}$
MIMO-VL 7B	48.1	76.7	24.7	58.3	52.0
+ SFT warm-up	49.2	78.0	43.8	69.0	60.0$_{+8.0}$
+ RL (DBN-RL)	45.9	78.8	45.2	71.9	**60.4**$_{+8.4}$

Implementation details. DBN-RL uses the same 3,000 CUB images as SpeciaRL for direct comparison; none of the evaluation datasets or DOWC-Taxo is used for training, and none overlaps in domain with CUB. We fine-tune the language tower with LoRA using the two-stage supervised warm-up and GRPO in Sec. 3, with $(\lambda_a, \lambda_b) = (0.1, 0.8)$. DBN-TF requires no training. We use $K = 3$ candidates for local discrimination and study K in Sec. 4.3. Training costs 5–7 GPU-hours per base model (Sec. 4.3). See App. E for details.

4.1 MAIN COMPARISON

Results on DOWC-Base. Tab. 1 shows a gap between text-guided and visual discrimination. Base models achieve around 90% on TC (fine-grained group), but only 3–48% and 10–29% on IV for the fine-grained and very fine-grained groups, respectively. DBN consistently improves discrimination across all five models. DBN-RL raises the discrimination average by 12.2 and 9.8 points on the two groups, while DBN-TF yields gains of 9.7 and 9.0 points; improvements are largest on the image-conditioned tasks (15–24 points). In contrast, SpeciaRL leaves discrimination nearly unchanged, indicating that naming-oriented training does not naturally induce discrimination. DBN also improves naming: DBN-RL gains 7.4 and 8.2 points, and DBN-TF gains 5.2 and 9.1 points. These results show that explicitly comparing local candidates benefits both discrimination and naming.

Results on DOWC-Taxo. Fig. 4 shows that base-model discrimination decreases monotonically as the taxonomic distinction becomes finer, falling from 42–52% at the family level to 30–39% at the species level. Averaged across ranks, DBN-RL and DBN-TF improve discrimination by 3.0 and 2.1 points, respectively. At the hardest species level, DBN-RL improves all five base models by 2.7–6.2 points. It also raises species-level naming by 3.2 points, compared with 2.7 points for DBN-TF and only 0.2 points for SpeciaRL (Tab. 2).

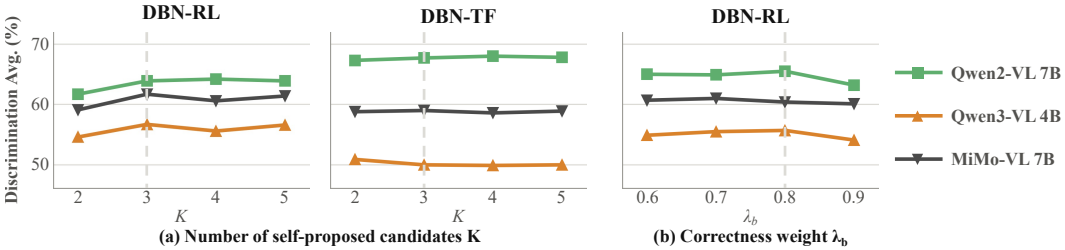


Figure 5: Hyper-parameter sensitivity on three models. (a) Effect of the number of self-proposed candidates K on DBN-RL and DBN-TF. (b) Effect of correctness-reward weight λ_B on DBN-RL.

4.2 Ablation Study

We ablate the two training stages of DBN-RL across five base models on the very fine-grained group (Tab. 3). The supervised warm-up teaches the model to propose and select among candidates before answering, while reinforcement learning further optimizes answer correctness and consistency with the selected candidate. Averaged over the five base models, the supervised warm-up improves discrimination by 6.9 points over the base, with the gains concentrated on IV and IC. Reinforcement learning provides an additional gain of 2.9 points over the warm-up and improves all five models.

4.3 Evaluation

Hyper-parameter sensitivity. We study the number of self-proposed candidates K and the DBN-RL reward weights. Fig. 5 (a) shows that increasing K from 1 to 2 improves discrimination by 2.1–2.6 points across three base models, while performance remains stable for $K = 3$ to 5. We therefore use $K = 3$, as larger values increase inference cost without clear gains. In DBN-RL, $\lambda_a = 0.1$ rewards well-formed responses, λ_b weights answer correctness, and $1 - \lambda_a - \lambda_b$ weights consistency with the selected candidate. Fig. 5 (b) shows stable performance for $\lambda_b = 0.6$ to 0.8. Removing the consistency term at $\lambda_b = 0.9$ reduces discrimination by 2.3, 1.6, and 0.3 points on the three models, respectively. We thus set $\lambda_b = 0.8$.

Cost. DBN-RL updates approximately 0.5% of the parameters using LoRA on a single GPU, requiring 5.3–7.2 GPU-hours per base model. It uses a rule-based reward and requires no judge model. In comparison, SpeciaRL fine-tunes all parameters on the same 3,000 images, requiring approximately 72 GPU-hours under its reported setting, and queries a 30B judge during training. DBN-TF requires no training but adds two short inference calls per image to construct the discrimination note. The note is cached and reused across questions, resulting in 5.8–6.7× the latency of direct inference on a single GPU. App. E provides a detailed cost breakdown.

5 Conclusion

We identified discrimination as a missing dimension of open-world classification: generating the correct name does not necessarily imply that an MLLM can distinguish the target from confusable alternatives. To measure this gap, we introduced DOWC, which jointly evaluates naming and discrimination, together with DOWC-Taxo, which controls difficulty through taxonomic distance. Our evaluation shows that discrimination remains particularly weak when models must rely on visual evidence alone, and that optimizing naming does not naturally improve this ability. We therefore proposed DBN, which explicitly compares self-proposed local candidates before answering. Across five base models, its training-based and training-free variants improve discrimination by 11.0 and 9.4 points and naming by 7.8 and 7.1 points, respectively. These results support a broader view of open-world recognition in which models must both name a class and reject plausible alternatives. Together, DOWC and DBN translate this broader view into a concrete benchmark and practical framework for evaluating and improving open-world models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and manuscript presentation, including improving the clarity, concision, grammar, and academic style of author-written text. The research questions, benchmark and method design, experiments, analyses, and scientific conclusions were developed and verified by the authors. All AI-assisted text was carefully reviewed and revised by the authors to ensure technical accuracy and consistency with the reported results. We take full responsibility for the final content of the paper.

REFERENCES

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 preprint arXiv:2404.14219*, 2024.

Moises Andrade, Joonhyuk Cha, Brandon Ho, Vriksha Srihari, Karmesh Yadav, and Zsolt Kira. Let's think in two steps: Mitigating agreement bias in mllms with self-grounded verification. In *International Conference on Learning Representations*, 2026.

Samuele Angheben, Davide Berasi, Alessandro Conti, Elisa Ricci, and Yiming Wang. Specificity-aware reinforcement learning for fine-grained open-world classification. *arXiv preprint arXiv:2603.03197*, 2026.

Niccolo Avogaro, Nayanika Debnath, Li Mi, Thomas Frick, Junling Wang, Zexue He, Hang Hua, Konrad Schindler, and Mattia Rigotti. Sparc: Separating perception and reasoning circuits for test-time scaling of vlms. *arXiv preprint arXiv:2602.06566*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Amartya Bhattacharya. Inference-time structural reasoning for compositional vision-language understanding. *arXiv preprint arXiv:2603.27349*, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024b.

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

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 2023.

Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Dmitry Demidov, Muhammad Zaigham Zaheer, Omkar Thawakar, Salman Khan, and Fahad Shahbaz Khan. Vocabulary-free fine-grained visual recognition via enriched contextually grounded vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

Dmitry Demidov, Muhammad Zaigham Zaheer, Zongyan Han, Omkar Thawakar, and Rao Anwer. Thinking beyond labels: Vocabulary-free fine-grained recognition using reasoning-augmented lmms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. Adapting vision-language models without labels: A comprehensive survey. *International Journal of Computer Vision*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, and Massimiliano Mancini. Compositional caching for training-free open-vocabulary attribute detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Marco Garosi, Matteo Farina, Alessandro Conti, Massimiliano Mancini, and Elisa Ricci. Large multimodal models as general in-context classifiers. *arXiv preprint arXiv:2602.23229*, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shaunak Halbe, Junjiao Tian, KJ Joseph, James Seale Smith, Katherine Stevo, Vineeth N Balasubramanian, and Zsolt Kira. Grounding descriptions in images informs zero-shot visual recognition. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. *arXiv preprint arXiv:2602.07605*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Gongyao Jiang and Qiong Luo. Chart-coca: Self-improving chart understanding of vision lms via code-driven synthesis and candidate-conditioned answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Hari Chandana Kuchibhotla, Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, and Vineeth N Balasubramanian. Efficient vocabulary-free fine-grained visual recognition in the age of multimodal llms. *arXiv preprint arXiv:2505.01064*, 2025.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 2024.

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 preprint arXiv:2408.03326*, 2024.

Jiaming Li, Zhijia Liang, Weikai Chen, Lin Ma, and Guanbin Li. Guided: Granular understanding via identification, detection, and discrimination for fine-grained open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 2025a.

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

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023b.

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

Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *International conference on learning representations*, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Quanyu Long, Kai Jie Jiang, Jianda Chen, Xu Guo, Leilei Gan, and Wenya Wang. Self-verification dilemma: Experience-driven suppression of overused checking in llm reasoning. *arXiv preprint arXiv:2602.03485*, 2026.

Barbara C Malt, Steven A Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40(2):230–262, 1999.

Georgiana Manolache, Gerard Schouten, and Joaquin Vanschoren. Crypticbio: A large multimodal dataset for visually confusing biodiversity. *arXiv preprint arXiv:2505.14707*, 2025.

Yuqi Peng, Pengfei Wang, Jianzhuang Liu, and Shifeng Chen. Glad: Generalizable tuning for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025.

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

Fawaz Sammani, Tzoulio Chamiti, and Nikos Deligiannis. On test-time scaling for vision-language models. In *European Conference on Computer Vision*, 2026.

Vishwas Sathish, Viresh Ranjan, Xinliang Zhu, Arnab Dhua, and Douglas Gray. Eliciting self-verification in multimodal reasoning agents with reinforcement learning. In *European Conference on Computer Vision*, pp. 384–401. Springer, 2026.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint arXiv:2603.06148*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sourabh Sharma, Sonam Gupta, and Sadbhawna. Improving reasoning in vision-language models via perception verified self-training. In *European Conference on Computer Vision*, 2026.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Catching the details: Self-distilled roi predictors for fine-grained mllm perception. In *International Conference on Learning Representations*, volume 2026, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good crepe needs more than just sugar: Investigating biases in compositional vision-language benchmarks. *arXiv preprint arXiv:2506.08227*, 2025.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

Madhukar Reddy Vongala, Saurabh Srivastava, and Jana Kosecka. Compositional image-text matching and retrieval by grounding entities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. Think or not? selective reasoning via reinforcement learning for vision-language models. *Advances in Neural Information Processing Systems*, 38:111196–111234, 2025a.

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 preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zilin Wang, Sangwoo Mo, Stella X Yu, Sima Behpour, and Liu Ren. Open ad-hoc categorization with contextualized feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Mingyuan Wu, Jingcheng Yang, Shengyi Qian, Xudong Wang, Jize Jiang, Qifan Wang, Aashu Singh, Khoi Pham, Fei Liu, Zhaolun Su, et al. Svr-r1: Bootstrapping multi-modal reasoning with self-verification in reinforcement learning. *arXiv preprint arXiv:2607.10966*, 2026.

Rui Xiao, Sanghwan Kim, Yongqin Xian, Zeynep Akata, and Stephan Alaniz. Finer: Mllms hallucinate under fine-grained negative queries. *arXiv preprint arXiv:2603.17662*, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models. *arXiv preprint arXiv:2502.09018*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jingxiao Yang, DaLin He, Miao Pan, Kaixiang Yao, Ge Su, Wenqi Zhang, Yifeng Hu, Tangwei Li, Yuke Li, and Xuhong Zhang. Sare: Sample-wise adaptive reasoning for training-free fine-grained visual recognition. *arXiv preprint arXiv:2603.17729*, 2026.

Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Towards open-ended visual recognition with large language models. In *European Conference on Computer Vision*, 2024.

Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. Object recognition as next token prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Fuxiang Zhang, Jiacheng Xu, Chaojie Wang, Ce Cui, Yang Liu, and Bo An. Incentivizing llms to self-verify their answers. *Advances in Neural Information Processing Systems*, 2025a.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025b.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 2022.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2025.

APPENDIX CONTENTS

A Related Work **16**

B Benchmark Details **17**

B.1 Prompt Templates 17

B.2 Auxiliary Metrics 18

B.3 Datasets and Statistics 19

B.4 DOWC-Taxo Construction 20

C DBN-RL Details **22**

C.1 Self-Generated Candidate Construction 22

C.2 Training Data 23

C.3 Training Details 23

D DBN-TF Details **24**

D.1 Prompts 24

D.2 Note Format and Note-Assisted Answering 24

E Additional Results **25**

E.1 Implementation Details 25

E.2 Cost Breakdown 25

E.3 Full Results on DOWC-Taxo 26

E.4 Per-Dataset Results on DOWC-Base 26

E.5 Qualitative Examples 26

A Related Work

Evaluating LMMs as image classifiers. Contrastive image–text pre-training turns class names into interchangeable text inputs (Radford et al., 2021) and, together with prompt learning (Zhou et al., 2022; Peng et al., 2025), supports open-vocabulary classification, yet a list of candidate names is still required at inference time. Vocabulary-free classification removes that list, either by retrieving candidates from a large text corpus (Conti et al., 2023; 2026) or by having the model generate the name directly (Yue et al., 2024; Yu et al., 2024; Huang et al., 2025). Large multimodal models have made the latter the default (Li et al., 2023a; Liu et al., 2023; Laurençon et al., 2024; Chen et al., 2024c; Wang et al., 2024; Bai et al., 2025b;a; Yang et al., 2025a; Yue et al., 2025). The setting has since been extended to few-shot adaptation, ad-hoc category discovery, grounded descriptions, attribute detection and label-free adaptation (Wang et al., 2025c; Halbe et al., 2026; Garosi et al., 2025; Dong et al., 2026), and some work queries the world knowledge of language models for class names directly (Liu et al., 2024). OWC (Conti et al., 2025) consolidates vocabulary-free classification into a unified evaluation of LMMs: it poses a single fixed open-ended question about the object in the image, and scores the emitted name with four complementary metrics—text inclusion, LLM judgment, semantic similarity and concept similarity—covering both string matching and semantics. Subsequent work continues to advance naming itself within this setting, either by changing the conditioning at inference time, using gallery exemplars as context and iteratively refining the prediction (Garosi et al., 2026) or generating candidate class names with an external language model or a reasoning LMM and then refining them (Demidov et al., 2025; 2026), or by changing the training signal, balancing the correctness and the specificity of the name with reinforcement learning (Angheben et al., 2026) and using an LMM to label unannotated images for weakly supervised fine-tuning (Kuchibhotla et al., 2025). All of this work scores recognition on a single axis: whether the emitted name is correct and specific enough. *In this paper, DOWC adds, alongside naming, a neglected but essential dimension of open-world classification—discrimination, that is, whether the model, given an image, chooses correctly between the ground-truth class and a visually confusable one. We cast discrimination into four question types that are scored by rule and that share their query images with naming, turning the evaluation of open-world classification from a single naming axis into two axes, naming and discrimination.*

Discriminative evaluation and hard negatives. Diagnosing LMMs with yes/no questions and paired scoring is by now standard practice, from hallucination benchmarks on object existence (Li et al., 2023b; Zhou et al., 2024), to pairing questions for consistency so that blind guessing no longer scores (Guan et al., 2024), to fine-grained negative queries and broad robustness suites (Xiao et al., 2026; Saxena et al., 2026). Constructing the negatives is the central issue of this line: negatives that carry exploitable distributional cues inflate the scores (Yuksekgonul et al., 2022; Udandarao et al., 2025), whereas adding hard negatives alone penalizes semantically equivalent hard positives (Kamath et al., 2024). The difficulty is drawn from different sources: disagreement between encoders (Tong et al., 2024), records of real-world confusion (Manolache et al., 2025), or generation by a language model (Zhang et al., 2025c); on the training side, similar but wrong categories serve as hard negatives or as region-level discrimination (He et al., 2025; Li et al., 2025a), and at inference time structured repair addresses failures of compositional understanding (Bhattacharya, 2026; Vongala et al., 2025). *In this paper, DOWC targets open-world classification and complements the accuracy of naming with a dimension of discrimination, where what is at stake is whether the name a model emits holds up against a visually neighboring class. The confusable class is given directly by the nearest class prototype under a frozen visual encoder (Siméoni et al., 2025), which needs no manually curated confusion list, no disagreement between two encoders and no generation by a language model, and therefore transfers automatically to any dataset with category labels.The positive instance is matched to the negative at the same level of similarity, so that difficulty leaks from neither side.*

Augmenting inference with candidates. Augmenting inference by proposing candidates has been realized in several ways in fine-grained recognition: the candidates are retrieved from an external database (Yang et al., 2026), drawn from the textual knowledge of a reasoning model or an external language model (Demidov et al., 2026; 2025), or generated once during training (Kuchibhotla et al., 2025), and chain-of-thought training, concept bottlenecks and expert collaboration realize the same pattern (He et al., 2026; Yamaguchi et al., 2025; Yang et al., 2025b). More generally, inserting an intermediate product before the answer covers free-form chains of thought and step-wise visual reasoning (Jaech et al., 2024; Xu et al., 2025; Thawakar et al., 2025; Zhang et al., 2025b), as well

as the debate over whether that reasoning has to be written out explicitly (Li et al., 2025b; Wang et al., 2025a); interventions at inference time that leave the parameters unchanged include test-time prompt tuning, test-time training and test-time scaling (Zuo et al., 2025; Sammani et al., 2026), together with plug-and-play pipelines such as perception–reasoning staging, caching and region proposal (Avogaro et al., 2026; Garosi et al., 2025; Bhattacharya, 2026; Shi et al., 2026; Dong et al., 2026). Closest in form to ours is the self-verification line: writing down the criteria before verifying against them, producing candidate answers before composing a final one, and eliciting self-verification or self-judged revision with reinforcement learning (Andrade et al., 2026; Jiang & Luo, 2025; Sathish et al., 2026; Wu et al., 2026; Zhang et al., 2025a; Sharma et al., 2026); excessive self-checking has also been reported to suppress performance (Long et al., 2026). *In this paper, we focus on open-world classification. Our DBN introduces no external retrieval database, no external language model and no additional classifier: the candidates are only the names the base model itself gives to the image at hand, and together they constitute a local discrimination. The model first makes a choice within this local discrimination and then answers conditioned on that choice, which brings into play a discriminative ability that free generation leaves unused, so that discrimination and naming improve together. We give two instantiations of the framework: DBN-TF constructs the local discrimination at inference time with the frozen base model and updates no parameters, whereas DBN-RL internalizes the same procedure into the model's own answering behavior through reinforcement learning with a rule-based reward.*

B Benchmark Details

B.1 Prompt Templates

Evaluation protocol. Every question in DOWC is posed in an independent single-turn conversation that contains only the image(s) and the question; no question is conditioned on the answers to other questions about the same image. All models are decoded greedily. The question templates are frozen and identical across models and datasets, so that differences between models are not confounded by prompt wording.

Naming. DOWC-Base uses the OWC naming prompt verbatim (Conti et al., 2025):

```
What type of object is in this photo?
```

DOWC-Taxo appends a suffix that fixes the target rank, since an iNat21 image admits correct answers at several taxonomic ranks:

```
What type of object is in this photo?  Answer only with the
name of its biological species.
```

The ground truth is the scientific name, and the iNat21 common name is accepted as an alias by the naming metrics.

Discrimination. Tab. 4 lists the four discrimination templates. The field {c} is a class name written exactly as in the dataset label set; {domain} is a single word fixed per dataset (cars, aircraft, flowers, pets, food for DOWC-Base; mammals, reptiles, or amphibians for DOWC-Taxo, according to the class of the query species). Each query image yields eight questions. TV asks the template twice, once with c and once with c^*. IV asks it twice, once with (x, x_c) and once with (x, x_{c^*}); the order in which the two images appear is fixed per question by a recorded random seed. TC and IC ask each question twice with the two candidate orders swapped, so that a fixed positional preference cannot score above chance.

Rank-aware variants for DOWC-Taxo. At the species rank, DOWC-Taxo uses the templates in Tab. 4 verbatim, with {c} set to the iNat21 common name. At the genus and family ranks, {c} is the scientific name of the taxon, and a positive pair in IV and IC may consist of two different species of the same genus or family, so the phrase *fine-grained category* would contradict the ground truth. The two image-based templates therefore replace it by the rank word, with {rank} $\in$ {genus, family}:

```
Do these two images belong to the same biological {rank} of
{domain}?  Answer yes or no.
```

Table 4: Frozen question templates of the four discrimination tasks. TV and IV are asked separately for the ground-truth class and the confusable class; TC and IC are asked in both candidate orders.

Task	Input	Question template	Answers
TV	$\langle x, c\rangle, \langle x, c^*\rangle$	`Is this a {c}? Answer yes or no.`	yes / no
IV	$\langle x, x_c\rangle, \langle x, x_{c^*}\rangle$	`Do these two images belong to the same fine-grained category of {domain}? Answer yes or no.`	yes / no
TC	$\langle x, c, c^*\rangle$	`Which of the following is the object in this image? A) {c} B) {c*}. Answer with the letter only.`	A / B
IC	$\langle x, x_c, x_{c^*}\rangle$	`The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.`	2 / 3

```
The first image is the query. Which of the other two images
shows the same biological {rank} as the query, the second or
the third? Answer 2 or 3.
```

TV and TC carry the class name explicitly and are unchanged across ranks.

B.2 Auxiliary Metrics

The main text reports one accuracy per discrimination task. This section defines those accuracies precisely, states their chance levels, and introduces two auxiliary quantities that we use to diagnose *why* a model scores as it does: the accept and reject rates of the verification tasks and the swap consistency of the choice tasks.

Verification tasks (TV, IV). For a query x with ground-truth class c and confusable class c^*, a verification task asks two questions, one about c (or x_c) and one about c^* (or x_{c^*}). Let a^+ and a^- denote the two parsed answers. The reported metric is the *paired accuracy*, $\mathbf{1}[a^+ = \text{yes} \wedge a^- = \text{no}]$, averaged over queries. A model that answers yes and no uniformly at random scores 25%, and a model that always accepts or always rejects scores 0. We additionally report the *accept rate*, the fraction of positive questions answered yes, and the *reject rate*, the fraction of negative questions answered no. The two rates locate a low paired accuracy on one side: Idefics2 8B, for instance, answers yes to 87% of positive and 73% of negative questions on Aircraft TV, so its paired accuracy of 18% reflects indiscriminate acceptance rather than a failure to recognize the true class.

Choice tasks (TC, IC). A choice task asks the same question twice with the two candidates swapped, and the reported *accuracy* is the fraction of the two answers that select the ground-truth class. Random guessing scores 50%, but so does a model that always selects the same position, since the ground truth occupies each position exactly once. To separate the two, we report the *swap consistency*, the fraction of queries for which both orders select the same class. A position-biased model has a swap consistency of 0, random guessing 50%, and a model that reads the image and ignores the order 100%. Phi-3-Vision on IC is the extreme case: it answers "2" on every question of all five datasets, giving an accuracy of exactly 50.0 with zero swap consistency.

Averages and chance levels. The Avg. column of each discrimination block is the unweighted mean of TV, TC, IV, and IC. Its chance level is therefore $(25 + 50 + 25 + 50)/4 = 37.5\%$, and the chance level of the verification tasks alone is 25%. Within DOWC-Base, fine-grained and very fine-grained scores are unweighted means over their three and two datasets, respectively.

Auxiliary metrics of the evaluated models. Tab. 5 reports the auxiliary metrics for the five base models used with DBN and for their DBN variants, averaged over the five DOWC-Base datasets with equal weight. Two patterns stand out. First, every base model is biased toward acceptance on IV: it answers yes to more than 86% of positive pairs but rejects only 6–46% of negative pairs, so the low IV scores in Tab. 1 reflect an inability to reject the confusable instance rather than to recognize the true one. Both DBN variants raise the reject rate substantially, in most cases to 60–80%, at the cost of some acceptance. Second, the base models' IC answers have swap consistencies of only 8–

66%, i.e., their choices depend heavily on the order in which the two instances are shown, whereas after DBN the consistency rises to 58–84%.

Table 5: Auxiliary discrimination metrics on DOWC-Base (in %, mean over the five datasets). Accept and reject rates are the fractions of positive and negative verification questions answered correctly; swap consistency is the fraction of queries for which both candidate orders select the same class.

	TV		IV		Swap consist.	
Model	accept	reject	accept	reject	TC	IC
Qwen2-VL 7B	95.3	58.5	99.6	6.0	85.9	8.0
+ DBN-TF	78.0	90.1	88.5	40.3	90.7	72.2
+ DBN-RL	73.3	85.1	77.2	74.4	86.6	69.9
Qwen2.5-VL 7B	85.2	69.9	94.5	36.7	90.3	65.8
+ DBN-TF	67.0	93.5	64.7	79.8	88.1	83.8
+ DBN-RL	78.3	82.1	68.8	76.3	90.2	69.8
Qwen3-VL 4B	86.3	68.9	93.9	37.1	88.1	34.9
+ DBN-TF	67.4	90.1	81.4	61.5	81.5	58.4
+ DBN-RL	62.6	87.5	73.8	72.9	84.3	76.6
Qwen3-VL 8B	81.6	75.6	93.4	46.0	90.5	47.9
+ DBN-TF	66.9	90.9	76.5	67.5	89.3	71.3
+ DBN-RL	70.0	84.9	63.0	75.4	84.3	76.9
MiMo-VL 7B	85.2	70.2	86.4	40.4	83.7	22.1
+ DBN-TF	61.3	93.0	72.8	65.9	79.1	77.5
+ DBN-RL	73.2	86.0	73.7	73.1	88.1	72.9

B.3 Datasets and Statistics

Sources and splits. DOWC-Base is built on the five fine-grained datasets of OWC and keeps the OWC train/val/test splits, which follow Conti et al. (2023). Each dataset plays two roles in DOWC. The *query* images are the test split: every query image is shown to the model and yields one naming question and eight discrimination questions (two per task, Sec. B.1). The *gallery* images are the union of the train and val splits: they are never shown as queries, but are used to build the questions, namely to compute the class prototypes μ_c that select the confusable class c^* and to provide the two instance images (x_c, x_{c^*}) of the image-based tasks (Sec. 2.2). Gallery and query images are disjoint, so an instance image is never a query image. All classes of the five datasets are kept. Tab. 6 summarizes the statistics.

Groups and averaging. Following OWC, Oxford Pets, Flowers102, and Food101 form the fine-grained (FG) group, and FGVC-Aircraft and Stanford Cars the very fine-grained (VFG) group. The datasets differ greatly in size: Food101 alone contributes 30,300 of the 47,806 query images (63%). Averaging over questions would therefore let one dataset dominate the FG group, so every group average and overall average in the paper is an unweighted mean over datasets, each dataset counting once.

How well the instance pairs are matched. The image-based tasks show the model a positive instance x_c (same class as the query) and a negative instance x_{c^*} (from the confusable class). If the negative instance were simply the gallery image of c^* most similar to the query and the positive instance an arbitrary image of c, the negative would usually look more like the query than the positive does, and the model could answer by similarity alone. Sec. 2.2 therefore picks the positive instance so that its similarity to the query matches that of the negative instance, s_{neg}. This matching is not always achievable: when s_{neg} is higher than the similarity of every gallery image of c to the query, the most similar one is taken; when it is lower than all of them, the least similar one is taken. We call these *endpoint* cases and record them. The last two columns of Tab. 6 report, per dataset, the percentage of endpoint cases and the residual similarity gap $|\text{sim}(x, x_c) - \text{sim}(x, x_{c^*})|$ averaged over queries. On Oxford Pets, Food101, Stanford Cars, and FGVC-Aircraft, endpoint cases are 5–23% of queries and the residual gap is at most 0.013, so positive and negative instances are almost equally

Table 6: Statistics of DOWC-Base. *Group*: OWC group of the dataset (FG: fine-grained; VFG: very fine-grained). *Classes*: number of classes. *Gallery*: images of the OWC train and val splits, used only to build questions (class prototypes and instance images). *Queries*: images of the OWC test split, each of which is the image the model is asked about. *Naming Q.*: naming questions, one per query image. *Discr. Q.*: discrimination questions, eight per query image. *Endpoint*: percentage of queries whose positive instance had to be taken at the boundary of the within-class similarity range because no gallery image of the true class was as similar (or as dissimilar) to the query as the negative instance (see text). $|\Delta\mathrm{sim}|$: mean absolute difference, after matching, between the query's cosine similarity to its positive instance and to its negative instance; smaller values mean the two instances are more tightly matched.

Dataset	Group	Classes	Gallery	Queries	Naming Q.	Discr. Q.	Endpoint (%)	$\lvert\Delta\mathrm{sim}\rvert$
Oxford Pets	FG	37	3,680	3,669	3,669	29,352	4.7	0.005
Flowers102	FG	102	5,726	2,463	2,463	19,704	51.5	0.064
Food101	FG	101	70,700	30,300	30,300	242,400	5.0	0.003
FGVC-Aircraft	VFG	100	6,667	3,333	3,333	26,664	22.8	0.013
Stanford Cars	VFG	196	8,144	8,041	8,041	64,328	7.5	0.012
Total		536	94,917	47,806	47,806	382,448		

similar to the query. Flowers102 is less tightly matched: about half of its queries are endpoint cases and the residual gap is 0.064. Similarities are cosine similarities of DINOv3 features, as in Sec. 2.2.

B.4 DOWC-Taxo Construction

Source images. DOWC-Taxo uses the validation split of iNat21 (Van Horn et al., 2021), which provides exactly 10 images per species together with the full taxonomic label of each species. We take *all* species of three classes, Reptilia (313 species), Mammalia (246), and Amphibia (170), without any further selection, giving 729 species and 7,290 images. According to the iNat21 taxonomy these species belong to 391 genera and 126 families. Every one of the 7,290 images serves as a query image at every rank; unlike DOWC-Base, there is no separate gallery, and the query set itself provides the class prototypes and the instance images, with the query image excluded from its own candidates. Fig. 6 shows query images with the instance pairs they receive at each rank.

One label set per rank. For a rank $r \in \{\text{species}, \text{genus}, \text{family}\}$, the classes are the taxa of that rank, and the images of a taxon are the images of all its member species. The construction of Sec. 2.2 is then applied without change to this label set: the prototype of a taxon is the mean feature of its images, the confusable taxon c^* is the other taxon of the same rank with the most similar prototype, the negative instance is the image of c^* most similar to the query, and the positive instance is the image of the query's own taxon whose similarity to the query is closest to s_{neg}. The confusable taxon is chosen by visual similarity alone and is not required to share a parent with the true taxon. What rank r guarantees is that c and c^* share no taxon at rank r or below, so the taxonomic distance between the two classes grows with r while the images stay the same. Since the taxonomy is fixed by iNat21 rather than by the evaluated model, this gives a difficulty axis that is independent of the model under test.

Table 7: Statistics of DOWC-Taxo per rank. *Taxa*: number of classes at that rank. *Images per taxon*: minimum / median / maximum number of images of a class. *Sim. to c^**: mean cosine similarity between the prototypes of a class and of its confusable class; lower values mean visually more distinct pairs. *Endpoint* and $|\Delta\mathrm{sim}|$ are defined as in Tab. 6. *Cross-species*: percentage of the 7,290 positive instance pairs (x, x_c) in which the two images belong to different species. *Discr. Q.*: discrimination questions at that rank (eight per query image).

Rank	Taxa	Images per taxon	Sim. to c^*	Endpoint (%)	$\lvert\Delta\mathrm{sim}\rvert$	Cross-species (%)	Discr. Q.
Species	729	10 / 10 / 10	0.790	33.9	0.035	0.0	58,320
Genus	391	10 / 10 / 200	0.783	24.2	0.025	39.1	58,320
Family	126	10 / 20 / 830	0.750	14.4	0.014	72.8	58,320

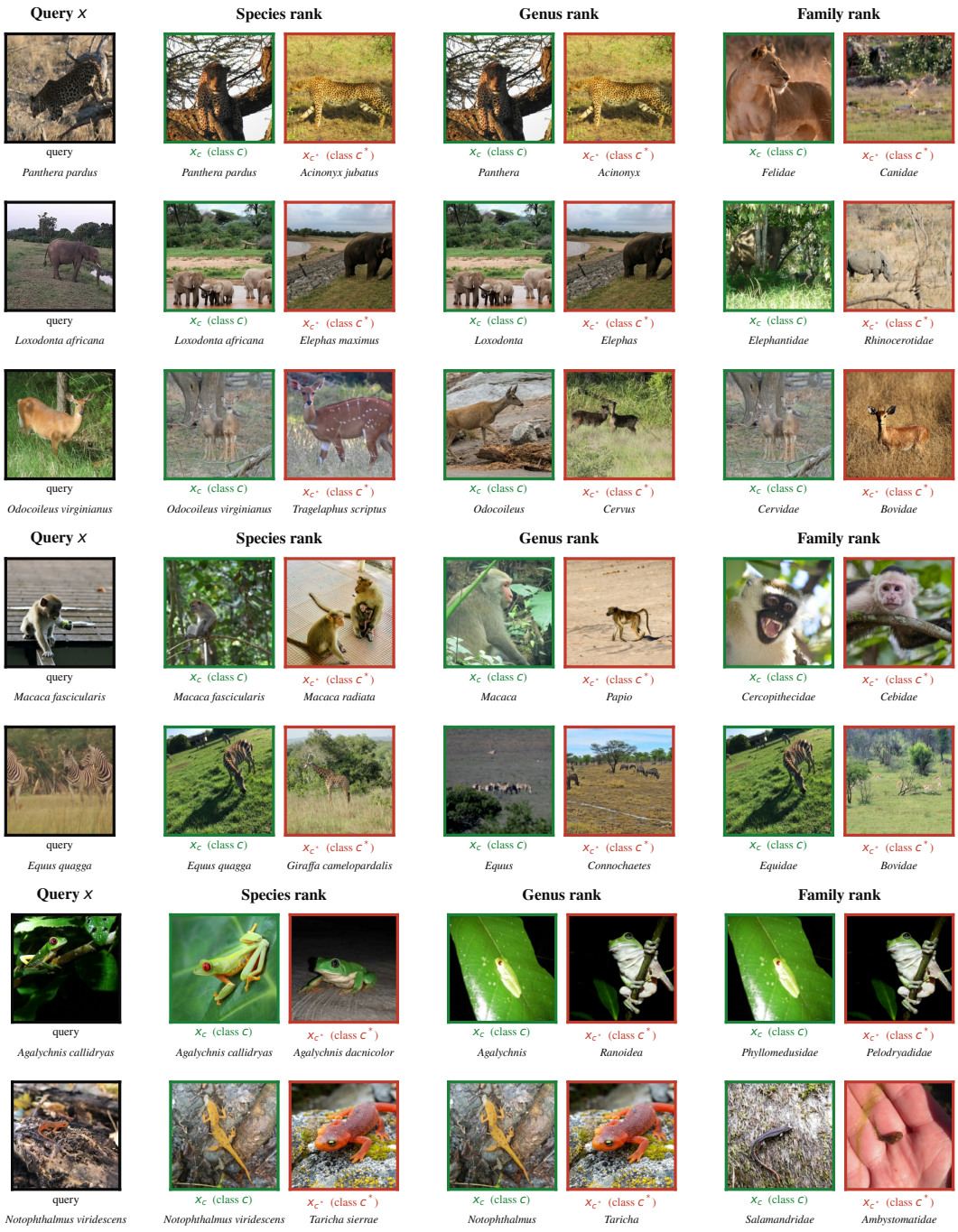


Figure 6: DOWC-Taxo examples. Each row is one query image (left) and, for the species, genus, and family ranks, the positive instance x_c (green, from the query's own taxon c) and the negative instance x_{c^*} (red, from the confusable taxon c^*) that the image-based tasks present; the names below the instances are the taxa c and c^* at that rank. As the rank coarsens, the confusable taxon moves from a sister species to another genus and another family (e.g., *Acinonyx jubatus*, *Acinonyx*, *Canidae* for the leopard), and the positive instance may come from a different species of the same taxon (e.g., a lion for the family Felidae). Because the confusable taxon is selected by visual similarity of prototypes rather than by the taxonomy, it can be taxonomically distant (a giraffe for a zebra at the species rank, a wildebeest at the genus rank). When a genus of the set contains a single species, the species- and genus-rank pairs coincide (e.g., *Loxodonta*). All instance pairs are taken verbatim from the benchmark.

Statistics. Tab. 7 summarizes the three ranks. As the rank coarsens, the prototype similarity between a class and its confusable class decreases from 0.790 to 0.750, i.e., the pairs become visually more distinct, which is the intended difficulty axis. Because a taxon at the genus or family rank pools several species, the positive instance of a query is frequently an image of a *different* species of the same taxon: this is the case for 39% of positive pairs at the genus rank and 73% at the family rank (52% of the family-rank positive pairs are even from different genera). The image-based questions must therefore ask about the taxonomic rank rather than about a "fine-grained category", which is why IV and IC use the rank-aware templates of Sec. B.1 at the genus and family ranks. The larger class pools also make similarity matching easier, so the endpoint rate and the residual gap decrease with rank. Each rank yields 58,320 discrimination questions, 174,960 in total.

Class names. At the species rank, TV and TC use the iNat21 common name of the species, which is available for all 729 species. At the genus and family ranks, iNat21 provides no common names, so TV and TC use the scientific name of the taxon (e.g., *Sciuridae*). The domain word of the image-based templates is set per class: `reptiles`, `mammals`, or `amphibians`.

Naming. Naming is evaluated once per image, at the species rank, giving 7,290 naming questions. The prompt asks for the biological species (Sec. B.1); the ground truth is the scientific name of the species, and its iNat21 common name is accepted as an alias by the naming metrics, so either name counts as correct.

C DBN-RL Details

This section gives the prompts and rules used to construct the self-generated candidates of Sec. 3.2 (Sec. C.1), the composition of the training data (Sec. C.2), and the training configuration (Sec. C.3).

C.1 Self-Generated Candidate Construction

All candidate names are produced by the base model θ_0 from the training images; no external model or class list is consulted. Three prompts are used. First, the coarse category of the image is obtained with

```
In one or two words, what general category of object is in
this photo?
```

and inserted as {`root`} into the following prompts. Second, the free-form naming of the image is obtained with

```
Name the {root} in this image as specifically as you can.
```

sampled several times at temperature 1; the names are normalized and the most frequent ones are kept. Third, confusable alternatives are obtained in a follow-up turn after the model's own greedy naming answer:

```
Give three other specific {root} models it could be confused
with, at the same level of specificity.  Answer with the
three names only, one per line.
```

Naming of the visually closest class. The prompts above only ask about the training image x itself. To make sure that the candidate set also contains the class that x is most likely to be confused with, we add one more name, obtained in two steps. In the first step we find the confusable class of x in the training set. This uses the same procedure as the benchmark construction in Sec. 2.2: every training class is represented by the mean DINOv3 feature of its images, and the confusable class c^* is the class whose prototype is closest to that of the ground-truth class of x. This step uses image features only; the base model is not involved. In the second step we obtain a name for c^*. We deliberately do not use the dataset label of c^*, because all candidates should be words the model produces by itself rather than labels copied from the dataset. Instead, we apply the free-form naming prompt above to the training images of c^*, collect the answers, and take the most frequent one as the name of c^*. The result is the name that θ_0 habitually gives to images of the neighboring class. It coincides with the dataset label of c^* when the model knows that class, and is the model's own (possibly incorrect)

name for it otherwise; in both cases it is a name that the model itself associates with images that look like x, which is what a confusable candidate should be.

Assembling C_x. The self-generated names are merged in the order (i) most frequent free-form name, (ii) name of the visually closest class, (iii) second most frequent free-form name, (iv) confusable alternatives, with normalization, de-duplication, and removal of generic words (e.g., `bird`, `object`). The first $K = 3$ names form $\mathcal{N}_{\theta_0}(x)$, the ground-truth class is added if absent, and the candidate order is shuffled, so that the ground truth occupies no fixed position. The same merging rule is used in the second warm-up stage, where the names come from θ_1's own notes instead of the probes above.

C.2 Training Data

Images and queries. Training uses the 3,000 CUB images of SpeciaRL (Angheben et al., 2026); no DOWC evaluation image, class, or domain is used. Discrimination queries for these images are built by applying the construction of Sec. 2.2 to the training images themselves, with the question templates of Sec. B.1, so that the training queries have the same form as the evaluation queries.

Records per stage. The first warm-up stage uses 3,000 naming records, one per image, whose target is the candidate note n_x^C. The second stage uses 27,000 records: 3,000 naming records and 6,000 records for each of TV, IV, TC, and IC, whose targets are the full note followed by the correct answer (Eq. (6)). For the multi-image queries, the note covers every image shown in the query and contains one selection per image. The reinforcement-learning stage samples from the same 27,000 queries without targets.

Target format. A note is a text block that lists the candidates and the selection, followed by the answer on a separate line. For a naming query, a second-stage target reads

```
<note>
domain: bird
candidates: Least Flycatcher, Pacific Wren, Pine Warbler,
Acadian Flycatcher
choice: Acadian Flycatcher
</note>
Answer: Acadian Flycatcher
```

where the domain line carries the coarse category, the candidates line C_x, and the choice line the selected candidate $\hat{c}(n)$. At inference, the model generates the whole block and the answer in a single turn, and only the text after `Answer:` is scored.

C.3 Training Details

Parameterization. All stages train LoRA adapters on the language model only: rank 16, scaling 32, no dropout, applied to the query, key, value, output, gate, up, and down projections of every decoder layer. The vision encoder, the vision–language projector, the token embeddings, and the output head are frozen. This amounts to about 0.5% of the parameters (40.4M for Qwen2-VL 7B and Qwen2.5-VL 7B, 33.0M for Qwen3-VL 4B, 43.6M for Qwen3-VL 8B, and 41.4M for MiMo-VL 7B). Training runs in bfloat16 on a single GPU.

Hyperparameters. Tab. 8 lists the settings of the three stages, which are identical for all base models. The second stage is initialized from the adapter of the first, and the reinforcement-learning stage from the adapter of the second, which also serves as the frozen reference policy for the KL term. Each reinforcement-learning step samples 16 queries (8 naming and 2 of each discrimination task) and $G = 8$ responses per query.

Validity and consistency rules. The indicator $v(o)$ in Eq. (7) is 1 only if the response contains one well-formed note block followed by exactly one `Answer:` line, the candidate list has no duplicates and stays within a length limit, and the answer is legal for the query (a non-empty name for naming, `yes`/`no` for TV and IV, `A`/`B` for TC, `2`/`3` for IC). Correctness compares the answer with the ground truth of the query; consistency compares the answer with the selection in the note as described in Sec. 3.2, with names matched after normalization. Both are computed by string rules without any model-based judge.

Table 8: Training configuration of DBN-RL (identical for all base models).

	Warm-up 1 ($\mathcal{L}_1$)	Warm-up 2 ($\mathcal{L}_2$)	GRPO
Records / queries	3,000	27,000	27,000
Steps	800	300	300
Batch size (per step)	32	32	16
Learning rate	1×10^{-4}	2×10^{-5}	1×10^{-5}
Optimizer	AdamW, wd 0.01, cosine schedule, 20 warm-up steps		
Group size G	–	–	8
Sampling	–	–	T=1.0, top-p 1.0, $\leq$160 new tokens
KL coefficient	–	–	0.01
Clipping ϵ	–	–	0.2
Reward weights (λ_a, λ_b)	–	–	(0.1, 0.8)

D DBN-TF Details

This section lists the prompts used by DBN-TF to construct the discrimination note (Sec. D.1) and the format in which the note is recorded and added to the queries (Sec. D.2). The base model is frozen throughout, all generations are greedy, and the evaluation questions are the templates of Sec. B.1.

D.1 Prompts

The three steps of Sec. 3.3 use two fixed prompts.

(1) Coarse-level scoping. The coarse category d_x is obtained with

```
Identify the main object in the image at a coarse category
level.  Answer with a short category name only.
```

The short answer (e.g., `bird`, `sports car`) is inserted into the next prompt.

(2) Fine-grained candidate proposal. Given the image and d_x, the K candidates C_x are obtained with the following prompt, where {`root`} is replaced by d_x ($K = 3$ in all experiments):

```
Name the {root} in this image as specifically as you can.
Give the 3 most likely answers.
Output exactly this and nothing else:
1.  <name>
2.  <name>
3.  <name>
```

The numbered lines of the response are parsed into the candidate list $C_x = (c_1, \ldots, c_K)$.

(3) Confidence-based selection. The highest-confidence candidate is selected as $\hat{c}_x$.

D.2 Note Format and Note-Assisted Answering

Note format. The discrimination note $n_x = \langle C_x, \hat{c}_x \rangle$ is recorded as a short text block that states the selection and lists the candidate neighborhood:

```
<note>
This object is {ĉ_x}.
candidates:  {c_1}, {c_2}, {c_3}
</note>
```

For example, for an image of a Pacific tree frog the note reads

```
<note>
This object is Pacific Tree Frog.
candidates:  Pacific Tree Frog, Northern Red-legged Frog,
California Red-legged Frog
```

```
</note>
```

Note-assisted answering. The note is generated once per image and cached. For every query about the image, the note is added to the query as additional context, introduced by a short fixed sentence that identifies it as the model's own discrimination analysis of the image; the question text itself is the unchanged template. For the multi-image tasks (IV and IC), each image shown in the question is accompanied by its own note, in the order in which the images appear, so the instance images x_c and x_{c^*} receive notes in the same way as the query image. The model then answers in a single turn, and its answer is scored exactly as for the base model.

E Additional Results

E.1 Implementation Details

Evaluation protocol. All models are evaluated with the OWC evaluation code, extended with the DOWC discrimination tasks. Decoding is greedy for every model and task, with a budget of 64 new tokens for naming questions and 8 new tokens for discrimination questions. For the Qwen-VL and MiMo-VL families, images are resized to at most $1024 \times 28 \times 28$ pixels; the other models use the preprocessing of the OWC code. MiMo-VL is run in its non-thinking mode by inserting an empty thinking block, both as a base model and with DBN. Naming metrics follow OWC: LI uses Llama-3.2-3B-Instruct with the OWC judge prompt, SS uses Sentence-BERT (all-MiniLM-L6-v2), and CS splits the prediction with spaCy (en_core_web_lg) before embedding. On DOWC-Taxo, where a species has a scientific and a common name, TI accepts either name, SS and CS take the maximum over the two, and LI receives both. Confusable pairs are constructed with DINOv3 ViT-B/16 features (L2-normalized CLS tokens), as described in Sec. 2.2.

Sources of the reported numbers. For the models of OWC and SpeciaRL, the naming numbers in Tab. 1 are taken from those papers, and all discrimination numbers are measured in this work. For the DBN rows and for the InternVL3.5, Qwen3-VL, and MiMo-VL base models, all numbers are measured in this work with the same code.

Evaluating DBN. DBN-RL generates the note and the answer in one pass; the generation budget is raised to 192 tokens so that the note fits, and only the text after the final `Answer:` marker is scored, with the same metrics and answer parsing as for the base model. DBN-TF adds the cached note to the input (Sec. D.2) and keeps the base model's generation budget. Every question is asked once; no sampling, voting, or retries are used.

E.2 Cost Breakdown

Training cost of DBN-RL. Tab. 9 lists the wall-clock time of the three training stages for each base model. Every stage runs on a single GPU; the two warm-up stages take about 1–1.5 hours in total and GRPO about 4–5.5 hours, giving 5.3–7.2 GPU-hours per base model. Peak memory stays below 21 GB in the warm-up stages and below 26 GB during GRPO (measured on MiMo-VL 7B, the largest run), so a single 40 GB GPU suffices. These figures cover the training loops; generating the candidate probes and the notes for the 3,000 training images adds about 2–4 GPU-hours per base model. For comparison, SpeciaRL reports reinforcement learning on the same 3,000 images with six A100 GPUs for approximately 12 hours (four for training and two for an online Qwen3-30B-A3B judge), i.e., about 72 GPU-hours, with full-parameter fine-tuning; this figure likewise excludes data preparation.

Inference cost of DBN-TF. DBN-TF adds two short generations per image, one for the coarse category (at most 8 new tokens) and one for the K candidates (at most $16K$ new tokens), which together produce 20–32 tokens per image on average. The note is cached per image and reused by every question in which the image appears, so the number of extra generations per question is two for the single-image tasks (naming, TV, TC), four for IV, and six for IC before caching, and less after. We measured latency on a single H100 with batch size 1 over 500 query images (250 each from Aircraft and Cars): direct answering takes 0.139 s per image for Qwen2-VL 7B and 0.158 s for MiMo-VL 7B, whereas the full DBN-TF pipeline takes 0.810 s and 1.058 s, i.e., 5.8× and 6.7×. About 71–73% of the DBN-TF time is spent generating the candidates; answering with the note in

Table 9: Training cost of DBN-RL per base model (single GPU). Minutes are training-loop wall-clock time of each stage; the total is in GPU-hours. Trainable parameters are those of the LoRA adapter.

Base model	GPU	Warm-up 1 (min)	Warm-up 2 (min)	GRPO (min)	Total (GPU-h)	LoRA params
Qwen2-VL 7B	A100 64GB	58.7	24.4	295.7	6.3	40.4M (0.49%)
Qwen2.5-VL 7B	H100	55.0	22.2	243.2	5.3	40.4M (0.49%)
Qwen3-VL 4B	H100 / A100	54.2	20.7	272.4	5.8	33.0M (0.74%)
Qwen3-VL 8B	H100	54.6	21.0	281.0	5.9	43.6M (0.50%)
MiMo-VL 7B	A100 64GB	77.5	31.2	326.2	7.2	41.4M (0.50%)

the input costs the same as answering without it, since the note adds only a few dozen text tokens to an image-dominated prompt.

E.3 Full Results on DOWC-Taxo

Tab. 10 reports the four discrimination tasks at each of the three ranks for the five base models with DBN-RL and DBN-TF, and for SpeciaRL on its Qwen2.5-VL 7B base; Fig. 4 in the main text plots the Avg. columns. Every entry is computed over the same 7,290 query images.

Table 10: Discrimination on DOWC-Taxo by rank (in %). Columns follow Tab. 1; Avg. is the mean of the four tasks (chance 37.5).

	Species					Genus					Family				
Model	TV	TC	IV	IC	Avg.	TV	TC	IV	IC	Avg.	TV	TC	IV	IC	Avg.
Qwen2-VL 7B	11.7	57.4	1.3	49.9	30.1	19.5	63.8	4.1	50.4	34.4	38.7	74.5	10.3	50.2	43.4
+ DBN-TF	20.6	57.2	16.5	53.4	36.9	30.9	60.8	18.2	54.7	41.1	48.8	74.1	22.0	58.4	50.8
+ DBN-RL	11.6	53.2	24.9	55.7	36.3	17.7	55.3	27.9	58.4	39.9	19.3	56.9	26.3	59.1	40.4
Qwen2.5-VL 7B	27.5	63.5	6.6	51.2	37.2	31.4	66.9	20.3	54.2	43.2	50.2	76.6	20.6	59.8	51.8
+ DBN-TF	19.4	54.2	20.7	55.1	37.4	35.6	58.3	27.4	59.5	45.2	48.0	73.3	29.9	65.8	54.2
+ DBN-RL	27.5	62.4	25.8	57.6	43.3	29.8	64.8	29.8	60.5	46.2	29.0	68.0	31.5	64.0	48.1
SpeciaRL (Qwen2.5-VL 7B)	27.1	63.5	6.4	51.0	37.0	27.8	66.5	19.1	53.0	41.6	49.0	76.9	20.9	60.0	51.7
Qwen3-VL 4B	20.8	62.1	2.2	51.9	34.3	23.6	67.7	28.7	51.3	42.8	34.3	78.8	32.0	48.0	48.3
+ DBN-TF	20.0	49.5	13.7	42.8	31.5	35.7	57.6	26.7	34.4	38.6	39.3	73.3	35.2	37.4	46.3
+ DBN-RL	13.0	59.6	26.6	58.4	39.4	24.7	62.4	28.1	62.1	44.3	42.9	65.8	25.3	65.7	49.9
Qwen3-VL 8B	24.5	61.7	13.5	54.3	38.5	30.0	67.9	25.2	55.1	44.6	51.5	79.6	20.2	57.5	52.2
+ DBN-TF	21.4	56.5	18.2	53.6	37.4	40.9	68.5	27.0	54.9	47.8	53.0	80.7	26.0	65.3	56.3
+ DBN-RL	16.0	61.3	28.4	59.1	41.2	25.0	67.2	32.2	62.1	46.6	35.9	69.5	35.7	66.9	52.0
MiMo-VL 7B	20.2	57.9	11.1	51.7	35.2	19.5	63.7	22.1	52.3	39.4	17.0	71.9	22.8	54.4	41.5
+ DBN-TF	18.6	47.9	16.0	54.5	34.2	31.5	58.0	20.2	57.1	41.7	43.9	67.8	22.9	64.0	49.6
+ DBN-RL	8.0	58.3	28.2	57.1	37.9	22.8	65.6	31.1	59.2	44.7	44.1	72.4	28.7	60.6	51.4

E.4 Per-Dataset Results on DOWC-Base

Tabs. 11 to 13 break the DOWC-Base results of Tab. 1 down by dataset for the five base models and their DBN variants. The group averages in Tab. 1 are the unweighted means of these per-dataset values.

E.5 Qualitative Examples

Figs. 7 to 16 show qualitative examples grouped by question type, two figures per type: the first covers Oxford Pets, Flowers102, and Food101, the second FGVC-Aircraft, Stanford Cars, and the species rank of DOWC-Taxo. Each example is a case in which the base model answers incorrectly and DBN-RL answers correctly. A card shows the image(s), the question, the ground truth, the base model's answer (red), the discrimination note generated by DBN-RL (its candidate list and selection), and its answer (green). All text is reproduced verbatim from the model outputs. For TV and IV the two answers are for the ground-truth and the confusable class (or instance), respectively; for TC and IC one of the two candidate orders is shown; for IV the note shown is the one written for

Table 11: Per-dataset results on DOWC-Base (in %) for the five base models and their DBN variants: Oxford Pets and Flowers102. Columns follow Tab. 1; Avg. is the mean of the four metrics of each block. Base-model naming is re-measured with our evaluation code for all five bases; Tab. 1 reports the published values for Qwen2-VL 7B and Qwen2.5-VL 7B, whose group averages differ from the re-measured ones by at most 1.4 points.

	Oxford Pets										*Flowers102*									
Model	TI	LI	SS	CS	Avg.	TV	TC	IV	IC	Avg.	TI	LI	SS	CS	Avg.	TV	TC	IV	IC	Avg.
Qwen2-VL 7B	12.3	48.7	25.1	43.3	32.4	52.1	91.8	0.2	50.4	48.6	42.1	72.4	41.9	69.0	56.3	58.5	89.6	2.0	56.1	51.5
+ DBN-TF	65.3	73.9	54.6	78.1	68.0	77.0	91.9	36.8	78.0	70.9	48.6	77.6	56.0	74.1	64.1	73.1	90.7	24.0	86.3	68.5
+ DBN-RL	71.9	75.7	82.2	83.6	78.3	77.9	91.9	62.2	83.4	78.8	46.8	66.3	66.7	67.4	61.8	56.7	85.1	64.3	87.6	73.4
Qwen2.5-VL 7B	65.3	71.9	41.2	77.3	63.9	55.3	92.5	8.8	53.1	52.4	56.4	75.2	47.4	79.7	64.7	67.8	92.2	71.9	74.4	76.6
+ DBN-TF	78.9	81.3	51.4	86.9	74.6	79.1	92.1	49.0	78.6	74.7	59.2	77.9	48.3	81.5	66.7	76.5	90.4	67.8	92.2	81.7
+ DBN-RL	77.7	79.1	84.4	85.9	81.8	72.5	93.0	59.9	81.8	76.8	43.8	67.9	62.9	64.1	59.7	69.3	92.7	68.2	90.5	80.2
Qwen3-VL 4B	68.5	74.7	42.4	73.4	64.8	60.4	89.0	1.2	56.8	51.8	51.8	72.5	45.5	71.6	60.4	66.3	88.4	78.8	75.5	77.3
+ DBN-TF	79.0	81.0	53.1	84.3	74.4	81.1	90.1	34.5	73.6	69.8	53.0	72.3	46.9	75.8	62.0	72.5	83.2	73.7	80.0	77.3
+ DBN-RL	78.6	79.8	82.7	85.9	81.7	79.2	89.6	56.9	83.0	77.2	37.6	60.2	61.1	61.9	55.2	62.6	88.6	72.4	93.1	79.2
Qwen3-VL 8B	69.5	75.0	43.3	75.1	65.7	65.2	90.4	18.0	61.4	58.7	54.0	71.2	46.1	72.2	60.9	71.6	92.7	82.8	81.7	82.2
+ DBN-TF	77.8	80.0	52.1	85.4	73.8	76.8	92.0	43.8	80.5	73.3	53.3	69.1	45.7	73.9	60.5	78.0	94.1	78.5	93.3	86.0
+ DBN-RL	73.9	76.8	84.0	84.9	79.9	75.3	90.4	57.0	79.2	75.5	37.2	65.6	56.8	59.2	54.7	69.1	90.7	60.4	94.9	78.8
MiMo-VL 7B	53.0	63.0	35.8	72.4	56.1	59.5	89.5	5.8	57.5	53.0	47.1	67.7	42.6	71.7	57.3	66.3	89.3	58.1	53.9	66.9
+ DBN-TF	67.1	73.8	52.7	80.7	68.6	71.6	88.7	26.7	77.5	66.1	42.2	77.3	40.2	70.3	57.5	67.5	86.9	61.5	86.1	75.5
+ DBN-RL	73.0	75.0	81.4	84.4	78.5	81.4	90.6	59.3	80.6	78.0	44.7	63.4	62.1	63.7	58.5	69.8	91.5	70.0	90.1	80.4

Table 12: Per-dataset results on DOWC-Base (in %): Food101 and FGVC-Aircraft, in the format of Tab. 11.

	Food101										*FGVC-Aircraft*									
Model	TI	LI	SS	CS	Avg.	TV	TC	IV	IC	Avg.	TI	LI	SS	CS	Avg.	TV	TC	IV	IC	Avg.
Qwen2-VL 7B	49.6	75.2	50.7	77.0	63.1	72.9	94.0	7.1	50.7	56.2	1.4	62.0	20.6	30.8	28.7	20.0	69.9	1.5	51.4	35.7
+ DBN-TF	45.1	58.1	56.3	71.8	57.8	74.9	92.6	32.2	71.1	67.7	52.6	71.6	49.6	73.1	61.7	66.1	78.9	37.5	72.1	63.6
+ DBN-RL	59.9	61.2	77.7	79.0	69.5	71.0	92.7	50.1	74.6	72.1	43.9	69.9	67.0	75.6	64.1	45.1	76.0	40.7	65.4	56.8
Qwen2.5-VL 7B	54.9	76.4	52.4	79.6	65.8	72.6	92.7	27.3	66.6	64.8	32.9	70.6	36.6	68.4	52.1	32.6	69.5	19.4	55.3	44.2
+ DBN-TF	54.5	69.3	50.1	79.0	63.2	66.8	84.7	34.7	71.9	64.5	53.2	64.5	45.6	78.5	60.4	56.3	72.9	41.0	69.5	59.9
+ DBN-RL	50.9	51.7	69.1	71.9	60.9	68.5	91.1	38.5	59.2	64.3	41.9	58.2	60.6	68.4	57.3	45.3	73.9	32.6	64.5	54.1
Qwen3-VL 4B	61.5	79.7	51.4	80.5	68.3	76.2	95.3	43.3	49.2	66.0	26.8	53.8	34.5	61.6	44.2	27.5	67.1	11.0	21.0	31.6
+ DBN-TF	62.4	68.4	53.4	81.7	66.5	75.3	88.0	51.3	65.4	70.0	28.3	54.7	37.1	65.1	46.3	36.9	52.5	24.6	40.6	38.7
+ DBN-RL	58.8	57.8	72.4	77.0	66.5	72.9	91.7	41.2	73.0	69.7	21.3	35.9	45.3	52.5	38.8	33.7	63.1	30.0	62.8	47.4
Qwen3-VL 8B	61.1	76.8	52.4	80.4	67.7	74.9	94.2	44.3	69.7	70.8	29.8	45.6	37.6	64.6	44.4	37.4	67.7	13.0	54.9	43.2
+ DBN-TF	59.1	63.0	54.4	79.6	64.0	71.7	89.7	49.1	78.3	72.2	31.9	49.8	38.0	66.5	46.5	37.5	62.5	23.3	59.9	45.8
+ DBN-RL	55.0	56.2	72.3	74.3	64.5	73.6	92.1	38.2	74.9	69.7	27.5	40.7	56.4	65.6	47.5	36.5	61.0	30.2	61.6	47.3
MiMo-VL 7B	48.2	70.9	48.8	76.2	61.0	69.6	92.8	35.5	59.1	64.3	17.0	53.7	29.4	56.0	39.0	33.2	64.8	13.5	55.9	41.8
+ DBN-TF	51.8	57.6	65.1	75.6	62.5	67.5	86.4	44.2	75.3	68.4	34.5	60.2	41.1	68.6	51.1	43.4	68.2	33.3	66.2	52.8
+ DBN-RL	56.7	57.7	74.4	76.0	66.2	75.6	92.8	42.8	70.6	70.5	27.6	49.4	58.4	65.9	50.3	44.3	68.4	33.4	62.9	52.2

Table 13: Per-dataset results on DOWC-Base (in %): Stanford Cars, in the format of Tab. 11.

	Stanford Cars									
Model	TI	LI	SS	CS	Avg.	TV	TC	IV	IC	Avg.
Qwen2-VL 7B	0.1	66.7	48.3	55.8	42.7	71.2	93.9	18.0	54.4	59.4
+ DBN-TF	0.0	84.4	63.9	72.8	55.3	72.5	93.9	36.9	84.3	71.9
+ DBN-RL	0.0	77.4	73.7	74.0	56.3	54.9	93.5	64.9	83.3	74.2
Qwen2.5-VL 7B	0.0	73.0	55.0	68.6	49.2	60.8	90.0	35.8	68.4	63.7
+ DBN-TF	0.0	71.2	56.9	73.6	50.4	45.0	89.4	53.1	80.8	67.1
+ DBN-RL	0.5	62.0	72.5	72.8	51.9	64.8	89.3	52.9	79.7	71.7
Qwen3-VL 4B	0.1	59.6	57.4	67.2	46.1	58.1	89.0	25.8	46.9	54.9
+ DBN-TF	0.1	74.5	56.6	69.6	50.2	48.3	79.3	48.8	66.7	60.8
+ DBN-RL	0.3	46.3	71.9	72.4	47.7	28.2	84.8	59.0	83.9	64.0
Qwen3-VL 8B	0.5	54.1	57.0	69.3	45.2	56.2	87.9	45.4	66.0	63.9
+ DBN-TF	0.1	69.8	56.1	69.3	48.8	49.3	84.5	47.9	78.2	65.0
+ DBN-RL	0.2	37.3	70.4	70.9	44.7	42.3	83.8	43.2	79.7	62.2
MiMo-VL 7B	0.0	60.8	45.1	55.8	40.5	63.0	88.6	35.9	60.7	62.0
+ DBN-TF	0.1	74.9	59.3	70.4	51.2	43.4	84.4	52.5	82.2	65.6
+ DBN-RL	0.0	76.3	70.7	71.0	54.5	47.5	89.1	57.0	80.8	68.6

the pair (query, x_{c^*}), with one selection per image. The DOWC-Base examples use Qwen2.5-VL 7B and the DOWC-Taxo examples Qwen3-VL 8B.

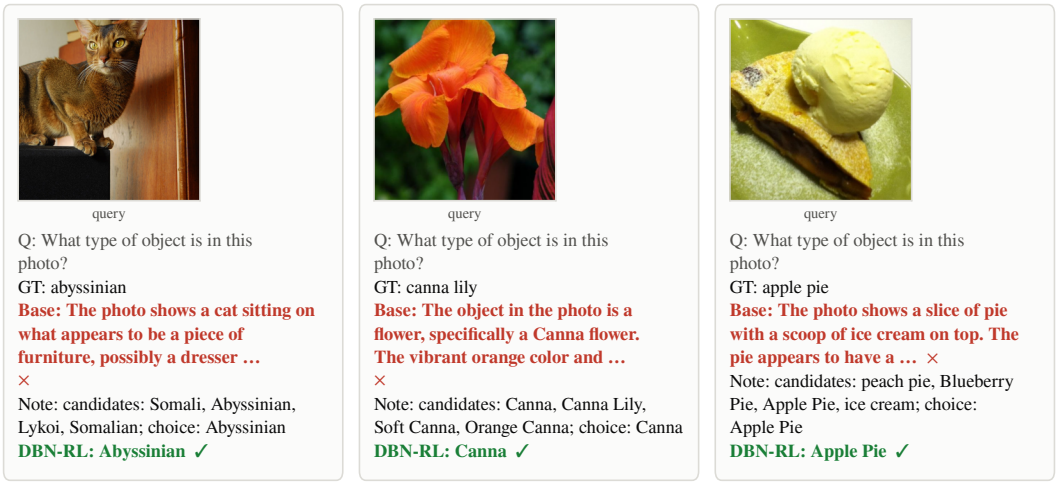


Figure 7: Qualitative examples for naming on Oxford Pets, Flowers102, and Food101 (Qwen2.5-VL 7B).

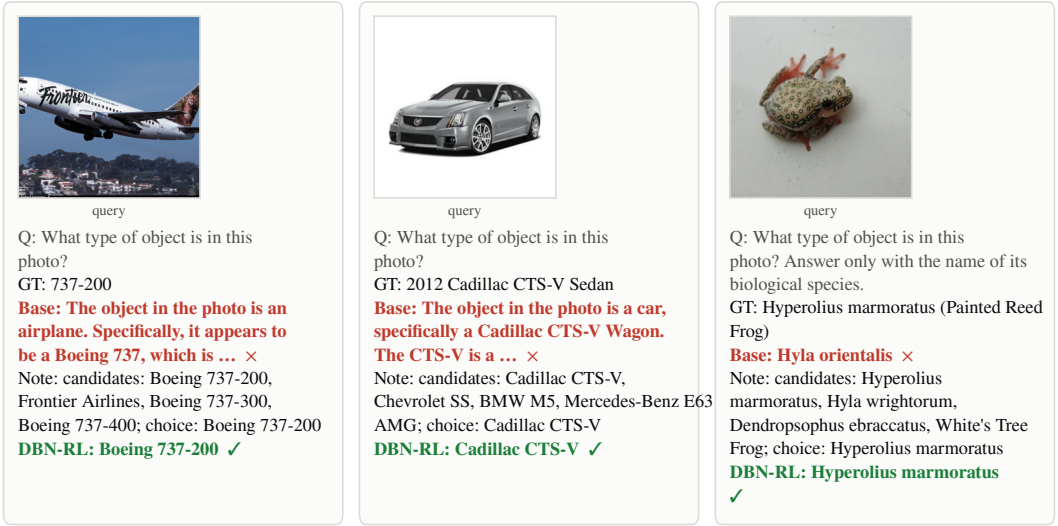


Figure 8: Qualitative examples for naming on FGVC-Aircraft, Stanford Cars, and DOWC-Taxo.



query

Q1: Is this a abyssinian? Answer yes or no.
Q2: Is this a bengal? Answer yes or no.
GT: yes / no
Base: yes / yes ×
Note: candidates: Somali, Abyssinian, American Bobtail, Jungle Cat; choice: Abyssinian
DBN-RL: yes / no ✓



query

Q1: Is this a water lily? Answer yes or no.
Q2: Is this a lotus? Answer yes or no.
GT: yes / no
Base: yes / yes ×
Note: candidates: Yellow Water Lily, American Lotus, European Water Lily, Nymphaea; choice: Yellow Water Lily
DBN-RL: yes / no ✓



query

Q1: Is this a apple pie? Answer yes or no.
Q2: Is this a bread pudding? Answer yes or no.
GT: yes / no
Base: no / no ×
Note: candidates: Lattice Pie, Deep-Dish Pizza, Apple Pie, Pot Pie; choice: Apple Pie
DBN-RL: yes / no ✓

Figure 9: Qualitative examples for Text-Verify (TV) on Oxford Pets, Flowers102, and Food101 (Qwen2.5-VL 7B).



query

Q1: Is this a 727-200? Answer yes or no.
Q2: Is this a DC-9-30? Answer yes or no.
GT: yes / no
Base: no / no ×
Note: candidates: Air Canada, Boeing 727-200, Canadian Airlines, Boeing 727; choice: Boeing 727-200
DBN-RL: yes / no ✓



query

Q1: Is this a 2012 GMC Terrain SUV? Answer yes or no.
Q2: Is this a 2012 GMC Acadia SUV? Answer yes or no.
GT: yes / no
Base: yes / yes ×
Note: candidates: GMC Terrain, 2012 GMC Terrain, 2013 GMC Terrain, 2011 GMC Terrain; choice: 2012 GMC Terrain
DBN-RL: yes / no ✓



query

Q1: Is this a Hyla versicolor? Answer yes or no.
Q2: Is this a Hyla chrysoscelis? Answer yes or no.
GT: yes / no
Base: yes / yes ×
Note: candidates: Pacific Tree Frog, Gray Tree Frog, Hyla versicolor, Squirrel Tree Frog; choice: Hyla versicolor
DBN-RL: yes / no ✓

Figure 10: Qualitative examples for Text-Verify (TV) on FGVC-Aircraft, Stanford Cars, and DOWC-Taxo.



query

Q: Which of the following is the object in this image? A) american bulldog B) american pit bull terrier. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: American Bulldog, Boxer, American Pit Bull Terrier, Bulldog; choice: American Bulldog
DBN-RL: A ✓



query

Q: Which of the following is the object in this image? A) water lily B) lotus. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: Pink Water Lily, American Lotus, European White Water Lily, Red Water Lily; choice: Pink Water Lily
DBN-RL: A ✓



query

Q: Which of the following is the object in this image? A) apple pie B) bread pudding. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: Apple Pie, Banana Split, Bread Pudding, ice cream; choice: Apple Pie
DBN-RL: A ✓

Figure 11: Qualitative examples for Text-Choose (TC) on Oxford Pets, Flowers102, and Food101 (Qwen2.5-VL 7B).



query

Q: Which of the following is the object in this image? A) 737-300 B) 737-500. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: Boeing 737-200, TAESA, Boeing 737-500, Boeing 737-300; choice: Boeing 737-300
DBN-RL: A ✓



query

Q: Which of the following is the object in this image? A) 2012 Hyundai Sonata Sedan B) 2012 Hyundai Azera Sedan. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: 2012 Hyundai Sonata, 2012 Hyundai Azera, 2013 Hyundai Sonata, 2012 Hyundai Genesis; choice: 2012 Hyundai Sonata
DBN-RL: A ✓



query

Q: Which of the following is the object in this image? A) Anaxyrus cognatus B) Anaxyrus woodhousii. Answer with the letter only.
GT: A
Base: B ×
Note: candidates: Colorado River Toad, Texas Toad, Great Plains Toad, Southern Toad; choice: Anaxyrus cognatus
DBN-RL: A ✓

Figure 12: Qualitative examples for Text-Choose (TC) on FGVC-Aircraft, Stanford Cars, and DOWC-Taxo.



query x_c x_{c^*}

Q: Do these two images belong to the same fine-grained category of pets? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}): yes / no
Base: yes / yes ×
Note for (query, x_{c^*}): candidates: Somali, Abyssinian, Ocicat, Bengal; choice 1: Abyssinian; choice 2: Bengal
DBN-RL: yes / no ✓



query x_c x_{c^*}

Q: Do these two images belong to the same fine-grained category of flowers? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}): yes / no
Base: yes / yes ×
Note for (query, x_{c^*}): candidates: Pink Water Lily, Sacred Lotus, European White Water Lily, American lotus; choice 1: Pink Water Lily; choice 2: Sacred Lotus
DBN-RL: yes / no ✓



query x_c x_{c^*}

Q: Do these two images belong to the same fine-grained category of food? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}): yes / no
Base: yes / yes ×
Note for (query, x_{c^*}): candidates: Ribs, Baked Potato, Barbecue ribs, steak; choice 1: Ribs; choice 2: steak
DBN-RL: yes / no ✓

Figure 13: Qualitative examples for Image-Verify (IV) on Oxford Pets, Flowers102, and Food101 (Qwen2.5-VL 7B).



Q: Do these two images belong to the same fine-grained category of aircraft? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}):
yes / no
Base: yes / yes ×
Note for (query, x_{c^*}):
candidates: Scandinavian Airlines, Boeing 737-600, Air Baltic, Boeing 737-500; choice 1: SAS Boeing 737-600; choice 2: Air Baltic Boeing 737-500
DBN-RL: yes / no ✓



Q: Do these two images belong to the same fine-grained category of cars? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}):
yes / no
Base: yes / yes ×
Note for (query, x_{c^*}):
candidates: Jeep Liberty, Jeep Patriot, Jeep Grand Cherokee, Jeep Commander; choice 1: Jeep Patriot; choice 2: Jeep Liberty
DBN-RL: yes / no ✓



Q: Do these two images belong to the same fine-grained category of mammals? Answer yes or no.
GT (query, x_c) / (query, x_{c^*}):
yes / no
Base: yes / yes ×
Note for (query, x_{c^*}):
candidates: Humpback Whale, Minke Whale, Blue Whale, Southern Right Whale; choice 1: Humpback Whale; choice 2: Minke Whale
DBN-RL: yes / no ✓

Figure 14: Qualitative examples for Image-Verify (IV) on FGVC-Aircraft, Stanford Cars, and DOWC-Taxo.



Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 2
Base: 3 ×
Note: candidates: Abyssinian, Somali, American Wirehair, Turkish Van; choice query: Abyssinian; choice second: Abyssinian; choice third: Bengal
DBN-RL: 2 ✓



Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 2
Base: 3 ×
Note: candidates: Cyclamen, sowbread, sowbread plant (Cyclamen persicum, Persian cyclamen; choice query: Cyclamen; choice second: Cyclamen; choice third: sweet pea
DBN-RL: 2 ✓

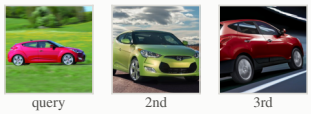


Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 2
Base: 3 ×
Note: candidates: Beef Carpaccio, Tuna Tartare, carpaccio, Beef Tartare; choice query: Beef Carpaccio; choice second: Beef Carpaccio; choice third: Tuna Tartare
DBN-RL: 2 ✓

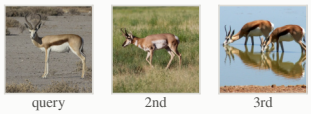
Figure 15: Qualitative examples for Image-Choose (IC) on Oxford Pets, Flowers102, and Food101 (Qwen2.5-VL 7B).



Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 2
Base: 3 ×
Note: candidates: TUI fly Boeing 737-800, Kelowna Flightcraft Boeing 737-800, Continental Airlines Boeing 737-900, Sunwing Airlines Boeing 737-800; choice query: TUI fly Boeing 737-800; choice second: Kelowna Flightcraft Boeing 737-800; choice third: Continental Airlines Boeing 737-900
DBN-RL: 2 ✓



Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 2
Base: 3 ×
Note: candidates: Hyundai Veloster, 2012 Hyundai Veloster, Hyundai Tucson, 2011 Hyundai Veloster; choice query: Hyundai Veloster; choice second: Hyundai Veloster; choice third: Hyundai Tucson
DBN-RL: 2 ✓



Q: The first image is the query. Which of the other two images shows the same fine-grained category as the query, the second or the third? Answer 2 or 3.
GT: 3
Base: 2 ×
Note: candidates: Springbok, Addax, Pronghorn, Scimitar-horned Oryx; choice query: Springbok; choice second: Pronghorn; choice third: Springbok
DBN-RL: 3 ✓

Figure 16: Qualitative examples for Image-Choose (IC) on FGVC-Aircraft, Stanford Cars, and DOWC-Taxo.