TARS: EXTENDING REINFORCEMENT LEARNING GAINS TO THE LONG TAIL UNDER LIMITED DATA FOR DOMAIN-SPECIFIC VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual recognition in domain-specific applications, such as biodiversity research and agriculture, often suffers from limited and imbalanced training data. Large Vision-Language Models (LVLMs) offer a promising starting point with rich visual and semantic knowledge acquired through web-scale pretraining, but naive supervised fine-tuning (SFT) on such data can cause them to overfit to head classes. Although reinforcement learning (RL) is reported to generalize better than SFT, our analysis shows that its gains under Group Relative Policy Optimization (GRPO) do not consistently extend to tail classes. To address this limitation, we propose **Tailness-aware Advantage ReScaling (TARS)**, a novel RL-based method that amplifies GRPO advantages for classes that are under-represented and poorly recognized. This amplification is guided by *tailness*, defined as the product of class rarity in the training data and the per-class error rate of the SFT model. Experiments on four domain-specific classification and detection benchmarks show that TARS consistently improves tail-class performance over SFT and vanilla GRPO, while outperforming existing imbalance-aware GRPO variants in both tail-class and overall performance. TARS also provides further gains when applied on top of imbalance-aware SFT methods, demonstrating that the two approaches are complementary.

1 INTRODUCTION

Visual recognition in domain-specific applications, such as classifying species for biodiversity research (Wah et al., 2011; Van Horn et al., 2018) and locating and identifying pests in agriculture (Wu et al., 2019), often suffers from limited data and class imbalance. Annotation requires domain expertise, which makes collecting large annotated datasets difficult. Moreover, some classes are encountered far more frequently than others during data collection in real-world settings, resulting in imbalanced training sets. Under these constraints, the objective is to fully leverage the scarce data while achieving balanced performance across classes along with high overall performance.

Recent Large Vision-Language Models (LVLMs) have shown strong capabilities after task-specific fine-tuning (Liu et al., 2024a; Bai et al., 2025a). Research on these models extends beyond multimodal reasoning and comprehension (Yue et al., 2024; Liu et al., 2024b) to visual recognition tasks such as classification and detection (Zhang et al., 2024; Yin et al., 2025; Conti et al., 2025; Li et al., 2025; Zang et al., 2025; Kisel et al., 2026). Though they may fail to recognize visual concepts in specific domains out of the box, their web-scale pretraining provides rich visual and semantic knowledge that is difficult to acquire by training from scratch on a small dataset. Thus, fine-tuning an LVLM is a promising starting point for domain-specific visual recognition under limited data.

However, supervised fine-tuning (SFT) learns by imitating target responses and can overfit to head classes under limited and imbalanced training data. Reinforcement learning (RL) allows models to explore responses guided by well-designed rewards and has been reported to generalize better than SFT (Chu et al., 2025), making it a promising approach to further improving generalization after SFT. A representative method, Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has been successfully applied to visual recognition with LVLMs (Liu et al., 2025). These findings suggest that GRPO may help models generalize from the few available examples of a tail class to

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

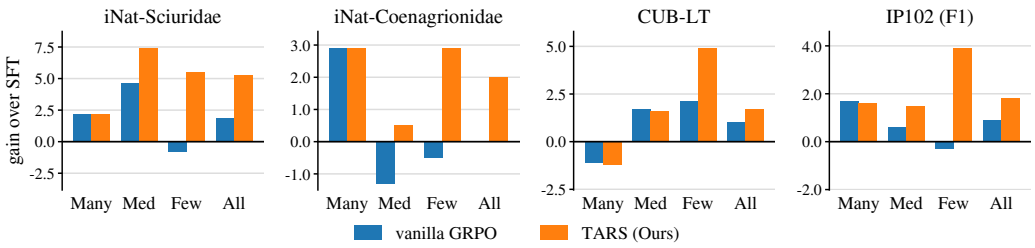


Figure 1: Gain over the CE loss SFT baseline, broken down by class frequency in the training data. Vanilla GRPO does not consistently improve performance on tail classes and sometimes underperforms SFT, while TARS achieves higher tail-class and overall performance than both SFT and vanilla GRPO on every benchmark. IP102 is reported in F1@0.5 and the others in accuracy.

its unseen instances. Yet, as shown in Fig. 1, vanilla GRPO does not consistently improve tail-class performance and even falls below the SFT baseline in several cases.

To understand this limitation, it is useful to consider two properties of GRPO. First, although the policy explores alternative responses, the objective of GRPO is still an expectation over queries drawn from the training data, so frequent classes contribute to policy updates more often than rare classes. Second, GRPO computes advantages by normalizing rewards across rollouts of the same query. This within-query normalization does not explicitly account for class frequency or how well the model recognizes the class. This motivates identifying classes with both limited exposure and substantial recognition error and amplifying the learning signals they produce.

To this end, we propose **Tailness-aware Advantage ReScaling (TARS)**, a novel RL-based generalization method that rescales GRPO advantages based on per-class *tailness* for domain-specific visual recognition. Tailness is a measure designed to identify classes that need stronger learning signals. Class frequency is a natural starting point, but frequency alone may not reflect how well the model recognizes each class (Sinha & Ohashi, 2023; Wang et al., 2021). For example, a tail class with distinctive visual features may already be well recognized. We therefore define tailness as the product of class rarity in the training data and the per-class error rate of the SFT model, accounting for both under-representation and recognition difficulty. TARS then selectively amplifies GRPO advantages based on class tailness. The same formulation applies across visual recognition tasks, including classification and detection.

We evaluate TARS on four domain-specific benchmarks with limited and imbalanced training data: iNat-Sciuridae and iNat-Coenagrionidae (Van Horn et al., 2018) and CUB-LT (Wah et al., 2011; Samuel et al., 2021) for classification, and IP102 (Wu et al., 2019) for detection. TARS consistently improves tail-class performance over SFT and vanilla GRPO, and outperforms imbalance-aware GRPO variants (Zhou et al., 2025; Dai et al., 2026a) in both tail-class and overall performance. Applying TARS on top of imbalance-aware SFT methods (Lin et al., 2017; Du et al., 2023; Kang et al., 2020) yields further gains, demonstrating their complementarity.

Our contributions are summarized as follows:

- We show that the generalization gains of vanilla GRPO do not consistently extend to tail classes in domain-specific visual recognition under limited and imbalanced training data.
- We propose **Tailness-aware Advantage ReScaling (TARS)**, a novel RL-based generalization method that selectively amplifies GRPO advantages using tailness, defined as the product of class rarity and the per-class error rate of the SFT model.
- Across four domain-specific classification and detection benchmarks, TARS improves tail-class performance over SFT and vanilla GRPO and outperforms imbalance-aware GRPO variants in both tail-class and overall performance. Further gains on top of imbalance-aware SFT methods demonstrate their complementarity with TARS.

2 Related Works

2.1 Visual Recognition with LVLMs

Recent studies directly evaluate the visual recognition capabilities of LVLMs and identify limitations in their out-of-the-box performance (Zhang et al., 2024; Kisel et al., 2026). Challenges include generating category names without a candidate list in open-world recognition (Conti et al., 2025) and connecting visual evidence to the correct concept name (He et al., 2025). To improve recognition, a line of work adapts LVLMs to classification and detection through supervised fine-tuning on task-formatted data (Yin et al., 2025; Li et al., 2025; Zang et al., 2025; Zhan et al., 2025). To address class imbalance during SFT, established methods from supervised recognition include loss reweighting (Lin et al., 2017; Du et al., 2023) and class-balanced sampling (Kang et al., 2020). Visual-RFT (Liu et al., 2025) applies GRPO to visual recognition using accuracy and IoU-based rewards, demonstrating gains in few-shot classification and detection. Subsequent work investigates RL for pure perception (Yu et al., 2026), combines detection, segmentation, and counting in a single model (Liu et al., 2026), and improves localization through RL-based post-training (Zhan et al., 2026; Jiang et al., 2026). TARS extends this line of work by allocating RL learning signals according to both class rarity and recognition error. It improves tail-class recognition under limited and imbalanced training data, while providing further gains on top of imbalance-aware SFT.

2.2 Reinforcement Learning for Generalization

RL with verifiable rewards has shown generalization benefits over SFT (Chu et al., 2025), motivating research on how to learn more effectively from problems that remain challenging for the model. In mathematical reasoning, GRPO-LEAD (Zhang & Zuo, 2025) and DGPO (Dai et al., 2026b) prioritize learning from questions with lower rollout success rates. Similarly, UniVG-R1 (Bai et al., 2025b) assesses the difficulty of individual visual grounding queries from their average grounding rewards across rollouts and prioritizes harder queries. Another line of work seeks to improve performance on under-represented domains during joint training. DISCO (Zhou et al., 2025) combines domain-frequency weighting with prompt-level difficulty across tasks such as mathematical reasoning, question answering, and sentiment classification. DRPO (Dai et al., 2026a) balances learning across clinical domains in multimodal medical reasoning based on domain frequency and reward statistics within each training batch. TARS seeks to extend RL's generalization gains to tail classes in domain-specific visual recognition under limited and imbalanced training data. It combines class rarity in the full training set with the SFT model's recognition error aggregated over the examples of each class. The resulting tailness determines how strongly GRPO advantages are amplified for each class.

3 Method

We consider domain-specific visual recognition over a set of classes $\mathcal{C}$, given a training set $\mathcal{D}$ that is limited in size and imbalanced over $\mathcal{C}$. Our RL stage starts from an SFT checkpoint and further optimizes the LVLM with GRPO, aiming for accurate recognition across all classes rather than only the frequent ones. We first review GRPO and discuss its limitations under class imbalance. We then introduce tailness and TARS, followed by the task-dependent verifiable rewards used in our experiments. Fig. 2 summarizes the resulting procedure.

3.1 Preliminaries

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) optimizes a policy without a learned value function by estimating advantages from the rewards of multiple responses to the same query. For a query q drawn from $\mathcal{D}$, the policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_i\}_{i=1}^{G}$, each of which receives a scalar reward r_i from a verifiable reward rule. The advantage of a response is obtained by standardizing the rewards within the group,

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G}) + \epsilon}, \tag{1}$$

where $\epsilon > 0$ is a tiny positive constant for numerical stability.

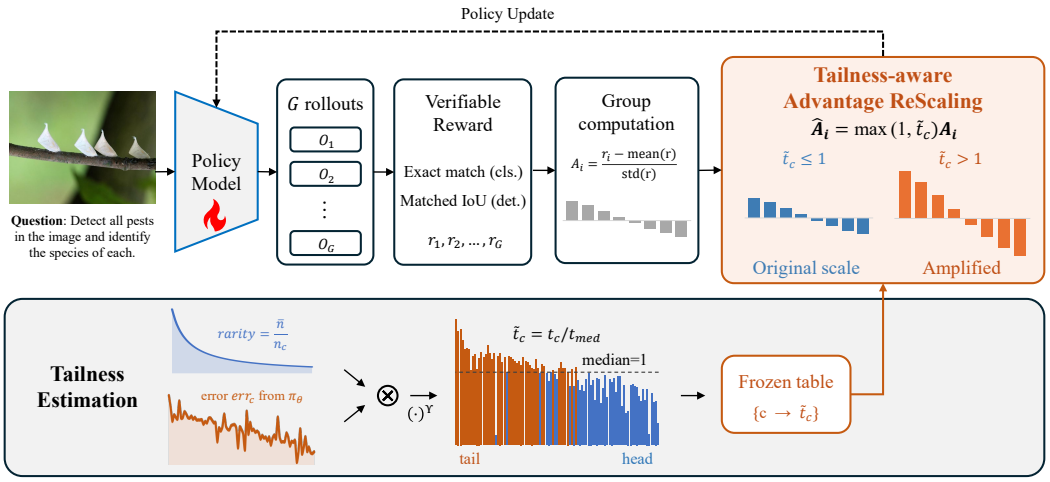


Figure 2: Overview of TARS. Class tailness is computed once before RL from class rarity and the SFT model's recognition error, then normalized by its median across classes. During RL, these fixed scores rescale GRPO advantages through $\max(1, \tilde{t}_c)$, amplifying signals for classes with above-median tailness while retaining the original scale for the rest.

For clarity, we present a simplified response-level formulation, omitting token-level details. For each response, GRPO forms the clipped surrogate (Schulman et al., 2017)

$$\ell_i(\theta) = \min\big(\rho_i A_i,\ \mathrm{clip}(\rho_i,\ 1-\varepsilon,\ 1+\varepsilon)A_i\big), \tag{2}$$

where $\rho_i = \pi_\theta(o_i \mid q)/\pi_{\theta_{\text{old}}}(o_i \mid q)$ is the importance ratio and ε sets the clipping range. Maximizing this surrogate encourages responses with positive advantages and discourages those with negative advantages. GRPO then maximizes

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\ell_i(\theta) - \beta D_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big]\right], \tag{3}$$

where the expectation is over $q \sim \mathcal{D}$ and $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$, and β weights the divergence from the reference policy π_{ref}.

Eq. 1 highlights a limitation central to this work. GRPO normalizes rewards using only the rollouts of the same query, without explicitly accounting for class frequency or aggregate class-level recognition performance. Thus, a poorly recognized tail class receives no additional amplification to compensate for its limited exposure. Meanwhile, frequent classes enter the objective more often because queries are drawn from $\mathcal{D}$, leaving the imbalance in class exposure uncorrected.

3.2 Tailness-aware Advantage ReScaling

Measuring tailness. We define tailness to identify classes that need stronger learning signals under imbalanced training data. Class rarity captures limited exposure, but a rare class may already be well recognized. We therefore combine rarity with the SFT model's per-class recognition error to account for how well each class is already learned.

Let n_c be the number of training queries in which class c occurs and let $\bar{n} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} n_c$ be the average occurrence count. Both classification and detection are posed as one query per image (see Tab. 1 for examples). Thus, n_c counts images labeled as c for classification and images containing at least one instance of c for detection. This measures class exposure at the query level rather than the instance level.

Before RL begins, we run a single inference pass over the training data with the initial SFT policy π_{θ_0} and measure its per-class performance $s_c \in [0, 1]$ using the same metric as at evaluation time.

Table 1: Query examples for classification and detection.

Task	Query prompt
Classification	What is the species of the *squirrel / damselfly / bird* shown in this image?
Detection	Detect all *pests* in the image and identify the species of each.

In our experiments, we use accuracy for classification and F1 score at an IoU threshold of 0.5 for detection. We define the recognition error as $err_c = 1 - s_c$ and the raw tailness of class c as

$$t_c = \left(\frac{\bar{n}}{n_c} \cdot err_c \right)^{\gamma}, \tag{4}$$

where $\bar{n}/n_c$ is the rarity of c relative to the average occurrence count and $\gamma > 0$ controls the contrast across classes. The product assigns greater tailness to rarer classes at the same error level, while reducing it for better-recognized classes at the same frequency. Setting $\gamma < 1$ compresses the relative differences between positive values, whereas $\gamma > 1$ accentuates them, in both cases preserving the ordering across classes.

Calibration. The scale of raw tailness varies across datasets, so we normalize it by the median tailness across classes,

$$\tilde{t}_c = \frac{t_c}{\max(t_{med}, \epsilon)}, \qquad t_{med} = \text{median}\big(\{t_c\}_{c \in \mathcal{C}}\big), \tag{5}$$

where $\epsilon > 0$ is a tiny positive constant to prevent division by zero. Thus, $\tilde{t}_c > 1$ indicates above-median tailness, reflecting both class rarity and recognition error. We use the median because it is less sensitive than the mean to extremely large raw tailness values.

Rescaling the advantage. For a query q associated with a single class $c(q)$, TARS multiplies its GRPO advantages by the class weight $w_c = \max(1, \tilde{t}_c)$ with $c = c(q)$. Since this weight is positive, scaling the advantages is equivalent to scaling the clipped surrogate in Eq. 2, yielding

$$\mathcal{J}_{\text{TARS}}(\theta) = \mathbb{E}_{q,\,\{o_i\}} \Big[\frac{1}{G} \sum_{i=1}^{G} \max\big(1,\ \tilde{t}_{c(q)}\big)\, \ell_i(\theta) - \beta\, D_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big] \Big]. \tag{6}$$

The scaling is one-sided. Classes with $\tilde{t}_c \leq 1$ retain the original GRPO advantage scale, while those with $\tilde{t}_c > 1$ receive amplification proportional to their calibrated tailness. This preserves the advantage signs and clipping conditions.

Tailness remains fixed throughout RL, so $\nabla_\theta[w_c \ell_i(\theta)] = w_c \nabla_\theta \ell_i(\theta)$. Each response's surrogate gradient therefore retains its direction, while reweighting the contributions of different classes can change the overall gradient direction. The KL term is not rescaled, reducing its relative influence for queries with larger weights.

Queries with multiple classes. In classification, each query corresponds to a single class. In detection, an image may contain multiple classes. Let $\mathcal{C}(q)$ denote the set of distinct classes present in the ground-truth annotations for query q. We average their weights,

$$w(q) = \frac{1}{|\mathcal{C}(q)|} \sum_{c \in \mathcal{C}(q)} \max\big(1, \tilde{t}_c\big), \tag{7}$$

and use $w(q)$ in place of $\max(1, \tilde{t}_{c(q)})$ in Eq. 6. This assigns a single amplification factor to the query based on the tailness of the multiple classes present in the image. The resulting query weight remains positive and fixed, preserving the scaling properties described above.

3.3 Verifiable Rewards for Visual Recognition

Following the use of verifiable rewards in Visual-RFT (Liu et al., 2025), we compute rollout rewards from classification correctness and, for detection, class-label agreement and bounding-box overlap. For both tasks, labels are converted to lowercase with separators removed before exact matching.

Classification. The model is asked to predict the class label of the image. The reward is an exact match between the normalized prediction $\hat{y}$ and ground-truth label y,

$$r = \mathbf{1}\big[\,\hat{y} = y\,\big]. \tag{8}$$

Any response that cannot be parsed into a class label receives zero.

Detection. The prompt asks the model to locate objects in the image and predict their class labels. The response is a JSON list in which each item contains a bounding box and a class label.

Let $\hat{\mathcal{B}} = \{(\hat{b}_k, \hat{y}_k)\}$ be the predicted set and $\mathcal{B} = \{(b_j, y_j)\}$ the annotated set. A predicted item is eligible to match an annotated item only if their normalized labels agree and their IoU is positive. Eligible pairs are matched greedily in descending order of IoU, with each item used at most once, yielding a matching $\mathcal{M}$. The reward is the total matched overlap divided by the larger of the two set sizes,

$$r = \frac{1}{\max\big(|\hat{\mathcal{B}}|,\, |\mathcal{B}|\big)} \sum_{(k,j)\in\mathcal{M}} \mathrm{IoU}\big(\hat{b}_k,\, b_j\big). \tag{9}$$

Matched boxes receive credit proportional to their IoU, while unmatched predictions and annotations remain included in the counts used for normalization. Unlike evaluation with F1 score at the IoU threshold of 0.5, the reward gives partial credit to matched boxes even below this threshold, reflecting differences in localization quality. A response that cannot be parsed into the required JSON form receives zero. We use no separate format reward.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TARS on four domain-specific benchmarks covering classification and detection. To study visual recognition under limited and imbalanced training data, we select benchmarks with at most 20,000 training images and an imbalance factor (IF) greater than 10. IF is the ratio of the training image count of the most frequent class to that of the rarest. For classification, we use two family-level subsets of iNaturalist-2018 (Van Horn et al., 2018), iNat-Sciuridae and iNat-Coenagrionidae, and CUB-LT (Samuel et al., 2021), derived from CUB-200-2011 (Wah et al., 2011), all from the biodiversity domain. For detection, we use IP102 (Wu et al., 2019) from agricultural pest management. iNat-Sciuridae has 4,996 training images from 36 squirrel species (IF 90.9), iNat-Coenagrionidae 4,042 from 62 damselfly species (IF 33.9), and CUB-LT 2,945 from 200 bird species (IF 14.3). IP102 has 15,178 training images covering 97 pest classes (IF 1,174) included in its bounding-box annotations.

Evaluation Protocol. All benchmarks use an open-world setting in which the model generates class names as free-form text without access to a candidate label list (Zhang et al., 2024). Following Liu et al. (2019), we divide classes by their training image count n_c into Many ($n_c > 100$), Medium ($20 \le n_c \le 100$), and Few ($n_c < 20$) for IP102 and the iNaturalist subsets. For CUB-LT, whose largest class has fewer than 100 training images, we use Many ($n_c > 20$), Medium ($5 \le n_c \le 20$), and Few ($n_c < 5$). For classification, we report accuracy based on exact matching between normalized predictions and ground-truth labels. For detection, the model predicts boxes and class labels without confidence scores, so we report precision, recall, and F1 score following Jiang et al. (2026). Within each image, predictions and ground-truth boxes of the same class are matched one-to-one using Hungarian matching at IoU ≥ 0.5. We compute these detection metrics per class and report the mean of each metric over classes with test annotations within each frequency group and across all classes.

Implementation Details. We adopt Qwen2.5-VL-3B-Instruct (Bai et al., 2025a) as our base LVLM and update all parameters except the vision encoder in both SFT and RL. Each GRPO optimization step uses 8 queries with group size $G = 8$, yielding a batch of 64 rollouts. We set the KL coefficient to $\beta = 0.04$ and the sampling temperature to 1.0. The tailness exponent is $\gamma = 0.75$ for classification and $\gamma = 1.0$ for detection. We repeat each RL run with three random seeds and report the mean. All experiments use bfloat16 precision on 4 NVIDIA H100 GPUs. Further implementation details are provided in Appendix A.

Table 2: Domain-specific classification accuracy (%) on iNat-Sciuridae, iNat-Coenagrionidae, and CUB-LT. Many, Medium, and Few group classes by training frequency, and All denotes overall accuracy. The first row shows Qwen2.5-VL-3B-Instruct without task-specific fine-tuning. SFT baselines use cross-entropy (CE Loss), focal loss (Focal Loss), class-frequency reweighting (Reweighting), or class-balanced sampling (Sampling). Bold and underlined values mark the best and second-best results in each column within each SFT block. Red subscripts show percentage-point gains over the corresponding SFT baseline for Few and All.

Method	iNat-Sciuridae				iNat-Coenagrionidae				CUB-LT			
	Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All
Qwen2.5-VL-3B	30.0	11.1	7.1	14.8	24.4	18.1	8.7	16.1	4.9	5.9	3.7	5.2
CE Loss	86.7	50.0	42.9	57.4	77.8	56.9	43.5	57.0	**92.7**	77.9	65.8	78.2
+ Vanilla	88.9	54.6	42.1	59.3	80.7	55.6	43.0	57.0	91.6	**79.6**	67.9	79.2
+ TARS (Ours)	**88.9**	**57.4**	**48.4**$_{\uparrow 5.5}$	**62.7**$_{\uparrow 5.3}$	**80.7**	**57.4**	**46.4**$_{\uparrow 2.9}$	**59.0**$_{\uparrow 2.0}$	91.5	79.5	**70.7**$_{\uparrow 4.9}$	**79.9**$_{\uparrow 1.7}$
Focal Loss	86.7	**58.3**	45.2	61.1	75.6	56.9	44.9	57.0	91.8	78.2	68.3	78.7
+ Vanilla	**90.0**	53.7	40.5	58.6	**77.8**	53.5	42.8	55.4	**92.3**	78.3	67.4	78.7
+ TARS (Ours)	88.9	57.4	**47.6**$_{\uparrow 2.4}$	**62.3**$_{\uparrow 1.2}$	76.7	**57.0**	**47.9**$_{\uparrow 3.0}$	**58.3**$_{\uparrow 1.3}$	92.0	**78.7**	**69.9**$_{\uparrow 1.6}$	**79.4**$_{\uparrow 0.7}$
Reweighting	86.7	58.3	45.2	61.1	77.8	**61.1**	44.9	59.1	91.7	78.6	69.2	79.1
+ Vanilla	90.0	58.3	43.6	61.4	82.2	56.9	43.0	57.9	**92.6**	79.2	67.0	79.0
+ TARS (Ours)	**91.1**	**62.9**	**49.2**$_{\uparrow 4.0}$	**65.4**$_{\uparrow 4.3}$	**82.2**	57.9	**48.3**$_{\uparrow 3.4}$	**60.2**$_{\uparrow 1.1}$	92.1	**79.7**	**72.0**$_{\uparrow 2.8}$	**80.4**$_{\uparrow 1.3}$
Sampling	**90.0**	58.3	54.8	65.7	73.3	52.8	43.5	54.3	89.6	**80.8**	73.9	81.0
+ Vanilla	90.0	61.1	53.2	66.1	74.1	56.0	44.9	56.3	89.8	79.2	72.6	79.8
+ TARS (Ours)	88.9	**62.0**	**60.3**$_{\uparrow 5.5}$	**68.8**$_{\uparrow 3.1}$	**75.6**	**56.5**	**51.7**$_{\uparrow 8.2}$	**59.3**$_{\uparrow 5.0}$	**90.1**	80.1	**76.0**$_{\uparrow 2.1}$	**81.3**$_{\uparrow 0.3}$

Table 3: Domain-specific detection results on IP102. We report precision, recall, and F1 score (%) at an IoU threshold of 0.5, averaged over classes within each frequency group and across all classes. SFT baselines and formatting follow Tab. 2. Red subscripts show TARS gains in percentage points over the corresponding SFT baseline for Few and All.

Method	P@0.5				R@0.5				F1@0.5			
	Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All
Qwen2.5-VL-3B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CE Loss	66.9	40.3	23.9	49.8	68.9	38.5	25.6	50.0	67.3	38.0	23.8	48.9
+ Vanilla	68.7	42.2	23.6	51.4	**70.3**	39.4	24.4	50.9	**69.0**	38.6	23.5	49.8
+ TARS (Ours)	**68.9**	**43.3**	**30.2**$_{\uparrow 6.3}$	**52.7**$_{\uparrow 2.9}$	70.1	**40.4**	**27.3**$_{\uparrow 1.7}$	**51.6**$_{\uparrow 1.6}$	68.9	**39.5**	**27.7**$_{\uparrow 3.9}$	**50.7**$_{\uparrow 1.8}$
Focal Loss	66.5	41.4	**39.4**	51.9	68.2	**41.2**	27.1	51.1	66.6	**39.7**	31.1	50.2
+ Vanilla	68.5	41.1	27.1	51.2	70.5	39.9	25.0	51.2	69.0	38.5	25.7	50.0
+ TARS (Ours)	**68.5**	**42.9**	37.8$_{\downarrow 1.6}$	**53.2**$_{\uparrow 1.3}$	**70.6**	40.2	**35.0**$_{\uparrow 7.9}$	**52.6**$_{\uparrow 1.5}$	**69.1**	38.9	**35.3**$_{\uparrow 4.2}$	**51.4**$_{\uparrow 1.2}$
Reweighting	66.7	44.0	32.7	52.4	69.1	**43.7**	27.1	52.7	67.2	**42.4**	28.4	51.4
+ Vanilla	**68.3**	43.1	37.3	53.2	**69.9**	42.1	36.8	**53.4**	**68.6**	41.1	35.9	**52.3**
+ TARS (Ours)	67.7	**44.4**	**40.8**$_{\uparrow 8.1}$	**53.9**$_{\uparrow 1.5}$	69.1	41.6	**38.9**$_{\uparrow 11.8}$	53.0$_{\uparrow 0.3}$	67.8	41.0	**39.0**$_{\uparrow 10.6}$	52.2$_{\uparrow 0.8}$
Sampling	63.0	**44.4**	39.7	51.8	65.6	**42.7**	39.2	52.1	63.6	**42.1**	37.3	50.7
+ Vanilla	62.6	43.5	41.3	51.4	66.0	40.6	36.2	51.0	63.8	39.7	37.2	49.7
+ TARS (Ours)	**63.0**	44.3	**44.9**$_{\uparrow 5.2}$	**52.3**$_{\uparrow 0.5}$	**66.6**	42.3	**40.8**$_{\uparrow 1.6}$	**52.5**$_{\uparrow 0.4}$	**64.0**	41.4	**40.8**$_{\uparrow 3.5}$	**51.0**$_{\uparrow 0.3}$

4.2 Main Results

We compare TARS with vanilla GRPO on top of four SFT methods, including standard cross-entropy loss, focal loss (Lin et al., 2017), class-frequency reweighting (Du et al., 2023), and class-balanced sampling (Kang et al., 2020). For each SFT method, TARS and vanilla GRPO start from the same checkpoint and use the same RL settings. Focal loss shifts learning emphasis toward harder predictions, while class-frequency reweighting and class-balanced sampling increase the contribution of under-represented classes through loss weighting and data sampling, respectively. These comparisons test whether TARS improves on vanilla GRPO across SFT initializations and provides further gains on top of imbalance-aware SFT.

Domain-specific Classification. Tab. 2 reports classification accuracy on the two iNaturalist subsets and CUB-LT, split by class frequency in the training set. The overall gains from vanilla GRPO do not consistently extend to tail classes, whose accuracy often drops below the SFT baseline. In

Table 4: Comparison with imbalance-aware GRPO variants using the same CE loss SFT checkpoint per dataset. We report accuracy (%) for classification and F1 (%) at IoU 0.5 for detection. Bold and underlined values mark the best and second-best results per column.

Method	iNat-Sciuridae				iNat-Coenagrionidae				CUB-LT				IP102			
	Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All
CE Loss	86.7	50.0	42.9	57.4	77.8	56.9	43.5	57.0	**92.7**	77.9	65.8	78.2	67.3	38.0	23.8	48.9
+ Vanilla	88.9	54.6	42.1	59.3	80.7	55.6	43.0	57.0	91.6	79.6	67.9	79.2	**69.0**	38.6	23.5	49.8
+ DISCO	87.8	57.4	43.7	60.5	**82.2**	55.6	44.9	58.1	91.5	79.0	68.5	79.0	65.8	37.1	26.3	48.2
+ DRPO	88.4	57.0	45.3	61.2	80.0	**57.7**	45.0	58.3	91.5	**79.8**	69.2	79.7	67.9	37.1	25.5	48.9
+ TARS (Ours)	**88.9**	**57.4**	**48.4**	**62.7**	80.7	57.4	**46.4**	**59.0**	91.5	79.5	**70.7**	**79.9**	68.9	**39.5**	**27.7**	**50.7**

contrast, TARS outperforms vanilla GRPO on both Few and All across all four SFT initializations on all three datasets. Starting from CE loss SFT, TARS improves Few accuracy over vanilla GRPO by 6.3, 3.4, and 2.8 percentage points on iNat-Sciuridae, iNat-Coenagrionidae, and CUB-LT, respectively, while also improving overall accuracy. Across all initializations, these gains come with at most a 1.1-point decrease in Many accuracy relative to vanilla GRPO. TARS also improves Few and All accuracy over every imbalance-aware SFT baseline. For example, on iNat-Coenagrionidae, applying TARS after class-balanced sampling raises Few accuracy from 43.5 to 51.7 and overall accuracy from 54.3 to 59.3. These results show that addressing imbalance during SFT does not eliminate the benefit of tailness-aware rescaling during RL.

Domain-specific Detection. Tab. 3 evaluates TARS on the detection benchmark IP102, complementing the classification results. Vanilla GRPO does not consistently improve tail-class performance. With SFT using CE and focal loss, it lowers Few F1 from 23.8 to 23.5 and from 31.1 to 25.7, respectively, while changing overall F1 by less than one percentage point. TARS instead raises Few F1 to 27.7 and 35.3, exceeding vanilla GRPO by 4.2 and 9.6 percentage points. Across all four SFT initializations, TARS outperforms vanilla GRPO on Few F1 and improves All F1 in three cases, with a 0.1-point decrease under class-frequency reweighting. TARS also improves Few and All F1 over all four SFT baselines. Together with the classification results, these findings demonstrate TARS's effectiveness across visual recognition tasks and its complementarity with imbalance-aware SFT.

Comparison with Imbalance-aware GRPO. We compare TARS with DISCO (Zhou et al., 2025) and DRPO (Dai et al., 2026a), which account for domain imbalance during RL. We adapt both methods by treating each class as a domain and initialize all methods from the same CE loss SFT checkpoint. As shown in Tab. 4, both methods improve Few performance over vanilla GRPO on all four benchmarks. TARS further outperforms both on Few and All across these benchmarks. On iNat-Sciuridae, it improves Few accuracy by 4.7 points over DISCO and 3.1 points over DRPO. On IP102, DISCO and DRPO improve Few F1 but reduce All F1 relative to vanilla GRPO, whereas TARS improves both. These results may reflect the granularity of the signal used to allocate RL updates. DISCO assesses prompt-level difficulty, while DRPO uses reward statistics within each batch. TARS instead combines class rarity with the SFT model's recognition error aggregated over all training data. This yields a stable class-level signal even when rare classes are absent from an individual batch.

Qualitative Results. Fig. 3 shows examples from Few classes across all four benchmarks. In the classification examples, the baselines misidentify tail classes as classes that occur frequently in the training data, while TARS predicts the correct labels. In the detection example, the baselines localize the objects but assign them to the frequent class *aphids*, whereas TARS correctly identifies the tail class *Toxoptera citricidus*. These cases illustrate the improvements in tail-class recognition observed in the quantitative results.

4.3 Ablation Studies

We conduct controlled ablations on iNat-Sciuridae, the most imbalanced classification benchmark, to evaluate the key design choices behind TARS. All variants start from the same SFT checkpoint and use the same RL configuration, changing only the component under study.

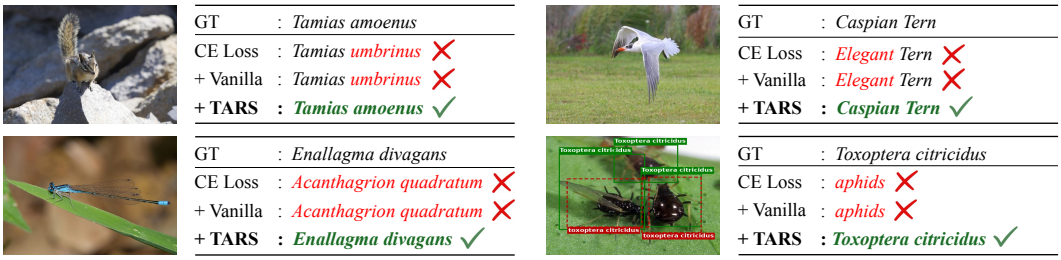
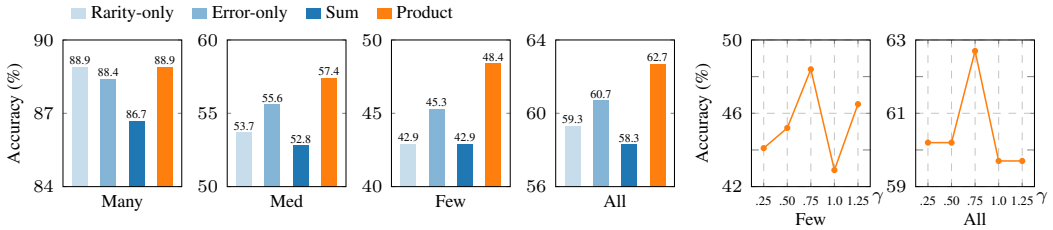


Figure 3: Qualitative results on Few classes across all four benchmarks. CE loss SFT and vanilla GRPO predict frequent classes, while TARS predicts the correct tail classes. For IP102, green boxes show ground truth and red boxes show predictions.



(a) Tailness design (b) Tailness sharpness γ

Figure 4: Ablations of tailness design and sharpness. (a) Accuracy with tailness based on rarity-only, error-only, their sum, or their product. (b) Few and All accuracy as the exponent γ varies.

Tailness Design. Fig. 4(a) compares the product of per-class rarity and recognition error against each signal alone and their sum, with the rest of TARS fixed. The product achieves the best Medium, Few, and All accuracy, reaching 48.4 on Few and 62.7 on All while matching rarity alone on Many. These results support using the product to account for limited exposure while tempering the emphasis on rare classes that are already well recognized.

Tailness Sharpness. The exponent γ controls the contrast among class weights, with larger values amplifying high-tailness classes more strongly. Fig. 4(b) shows that stronger amplification does not consistently improve performance. Among the tested values, $\gamma = 0.75$ achieves the highest Few and All accuracy, and we use it for classification.

Additional ablations in Appendix B show that one-sided rescaling with $\max(1, \tilde{t}_c)$ preserves performance in low-tailness classes while improving All accuracy, and that static tailness outperforms epoch-wise re-estimation on Few and All.

5 Conclusion

We studied domain-specific visual recognition with LVLMs under limited and imbalanced training data and showed that vanilla GRPO's gains do not consistently extend to tail classes. Its within-query advantage normalization does not explicitly account for class frequency or class-level recognition performance. We proposed TARS, which rescales GRPO advantages using tailness, a per-class measure combining class rarity with the SFT model's recognition error. Tailness is estimated once before RL and held fixed, requiring no additional inference for tailness estimation during training. Across four classification and detection benchmarks, TARS consistently improves tail-class performance over SFT and vanilla GRPO and outperforms the compared imbalance-aware GRPO variants in both tail-class and overall performance. It also provides further gains on top of imbalance-aware SFT methods, demonstrating their complementarity. The current formulation relies on class labels to estimate rarity and recognition error. Extending TARS to tasks without explicit classes would require alternative ways to measure exposure and model performance.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript. Additionally, we used generative AI tools for retrieval and discovery, such as identifying potentially relevant related work. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. In particular, all AI-assisted writing was reviewed and edited by the authors, and literature identified with AI assistance was independently verified against the original sources before being cited. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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. *arXiv preprint arXiv:2502.13923*, 2025a.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025b.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning*, pp. 10818–10838. PMLR, 2025.

Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16388–16398, 2025.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026a.

Yanqi Dai, Yuxiang Ji, Xiao Zhang, Yong Wang, Xiangxiang Chu, and Zhiwu Lu. Harder is better: Boosting mathematical reasoning via difficulty-aware grpo and multi-aspect question reformulation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=nfURupkdRJ.

Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15814–15823, 2023.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Nikita Kisel, Illia Volkov, Klara Janouskova, and Jiri Matas. Multimodal large language models as image classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1711–1720, June 2026.

Jincheng Li, Chunyu Xie, Ji Ao, Dawei Leng, and Yuhui Yin. Lmm-det: Make large multimodal models excel in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 308–318, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020.

Dvir Samuel, Yuval Atzmon, and Gal Chechik. From generalized zero-shot learning to long-tail with class descriptors. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 286–295, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Saptarshi Sinha and Hiroki Ohashi. Difficulty-net: Learning to predict difficulty for long-tailed recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6444–6453, 2023.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9695–9704, June 2021.

Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8787–8796, 2019.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14358–14368, 2025.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:94827–94853, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843, 2025.

Yufei Zhan, Shurong Zheng, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22947–22957, 2025.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5642–5665, 2025.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

Yuhang Zhou, Jing Zhu, Shengyi Qian, Zhuokai Zhao, Xiyao Wang, Xiaoyu Liu, Ming Li, Paiheng Xu, Wei Ai, and Furong Huang. DISCO balances the scales: Adaptive domain- and difficulty-aware reinforcement learning on imbalanced data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1407–1419, 2025.

A Additional Implementation Details

A.1 Training settings

We use AdamW (Loshchilov & Hutter, 2019) with a cosine learning-rate schedule and a warmup ratio of 0.03 for both SFT and RL. Tab. 5 lists the benchmark-specific settings, which are shared across methods within each benchmark. Images use a pixel budget ranging from $512 \times 28 \times 28$ to $1280 \times 28 \times 28$ during SFT, RL, and evaluation. SFT uses a global batch size of 64 examples, while each RL step uses 8 queries with 8 rollouts each. We train with DeepSpeed ZeRO-2 (Rajbhandari et al., 2020).

A.2 Imbalance-aware SFT Formulation for LVLMs

Cross-entropy, focal loss (Lin et al., 2017), and class-frequency reweighting (Du et al., 2023) differ in their loss weights, while class-balanced sampling (Kang et al., 2020) retains cross-entropy and changes the sampling distribution. The loss-based SFT baselines differ in how they weight answer tokens. Let $\mathcal{B}$ denote an SFT mini-batch, x_i the image and question of sample i, and y_i its ground-truth answer. We compute the loss over the T_i tokens in y_i. Let $p_{i,t} = \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$ denote the probability assigned to the t-th answer token given the input and preceding ground-truth tokens. The loss is

$$\mathcal{L} = -\frac{1}{\sum_{i\in\mathcal{B}} T_i} \sum_{i\in\mathcal{B}} \sum_{t=1}^{T_i} W_{i,t} \log p_{i,t}. \tag{10}$$

For the loss-based classification baselines, we instantiate the token weights in Eq. 10 as follows:

$$W_{i,t} = \begin{cases} 1, & \text{cross-entropy,} \\ (1-p_{i,t})^{\gamma_\mathrm{f}}, & \text{focal loss,} \\ w'_{c_i}, & \text{class-frequency reweighting,} \end{cases} \qquad w'_c = \frac{|\mathcal{C}|\, n_c^{-k}}{\sum_{j\in\mathcal{C}} n_j^{-k}}, \tag{11}$$

where n_c is the training image count defined in Eq. 4. Focal loss reduces the contribution of tokens that the model predicts with high confidence, while class-frequency reweighting gives greater weight to under-represented classes. The values of $\gamma\mathrm{f}$ and k are listed in Tab. 5.

Class-balanced sampling is a separate data-level baseline. It selects a class uniformly and then samples an image uniformly from that class for classification. Sampling uses replacement and retains the original number of training examples per epoch, keeping the number of optimization steps unchanged.

For detection, focal loss is applied to all answer tokens, including class labels, coordinates, and JSON structure. Class-frequency reweighting applies w'_c to tokens corresponding to the class label and bounding box of each object. Class-balanced sampling assigns each image a weight based on the inverse frequency of the rarest class it contains and samples images accordingly.

Table 5: Training settings for each benchmark. γ_f and k are the focal loss and class-frequency reweighting exponents, respectively.

Benchmark	SFT LR	SFT epochs	RL LR	RL epochs	γ_f	k
iNat-Sciuridae	5×10^{-6}	5	5×10^{-6}	3	1.0	0.25
iNat-Coenagrionidae	5×10^{-6}	5	5×10^{-6}	3	2.0	0.25
CUB-LT	5×10^{-6}	5	1×10^{-6}	3	1.0	0.25
IP102	5×10^{-5}	3	5×10^{-6}	1	1.0	0.25

B Additional Ablation Studies

B.1 One-sided Rescaling

We ablate the one-sided design of TARS by replacing $w_c = \max(1, \tilde{t}_c)$ with $w_c = \tilde{t}_c$, while keeping all other settings unchanged. This reduces the magnitude of advantages for classes with $\tilde{t}_c \leq 1$ below

the standard GRPO scale. Classes with $\tilde{t}_c > 1$ receive the same amplification in both variants. As shown in Fig. 5, removing one-sided rescaling changes the gain over the SFT checkpoint from $+1.2$ to -0.6 points for classes with $\tilde{t}_c \leq 1$. Performance also decreases for classes with $\tilde{t}_c > 1$, resulting in lower overall accuracy. These results support the one-sided design of TARS, which preserves the standard GRPO advantage scale for lower-tailness classes while selectively amplifying higher-tailness classes.

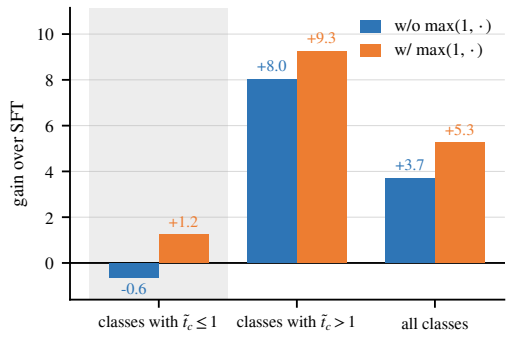


Figure 5: Effect of one-sided rescaling on iNat-Sciuridae. Accuracy gains over the CE loss SFT checkpoint in percentage points, grouped by calibrated tailness.

B.2 Static vs. Dynamic Tailness

TARS computes tailness once from the initial SFT policy and keeps it fixed throughout RL. To examine whether the error component should track the evolving policy, we compare this static design with an epoch-wise dynamic variant. At the end of each epoch, the dynamic variant runs greedy inference over the training set with the current policy, recomputes the per-class recognition errors, and updates the tailness weights for the following epoch. As shown in Tab. 6, dynamic re-estimation improves Many accuracy but lowers Few and All accuracy relative to static TARS. The dynamic variant also requires an additional 0.98% training time. These results support keeping tailness fixed throughout RL in our setting.

Table 6: Comparison between static and dynamic tailness on iNat-Sciuridae. We report accuracy (%) and additional training time relative to static TARS.

Tailness	Many	Medium	Few	All	Additional time
Dynamic	**90.0**	**57.4**	46.8	62.3	+0.98%
Static (Ours)	88.9	**57.4**	**48.4**	**62.7**	–

C Tailness Statistics

Fig. 6 visualizes the two factors of tailness and the resulting calibrated tailness on all four benchmarks. Rarity is inversely proportional to class frequency, so a weight built on it alone would reduce to frequency-based reweighting. However, the SFT model's per-class recognition error shows no strictly monotonic trend with class frequency. Thus, the product is used to capture both class rarity and recognition error, giving greater emphasis to classes that are under-represented and poorly recognized. Tailness therefore follows neither factor on its own, and the classes it lifts are the ones actually starved of signal, consistent with the tailness design ablation.

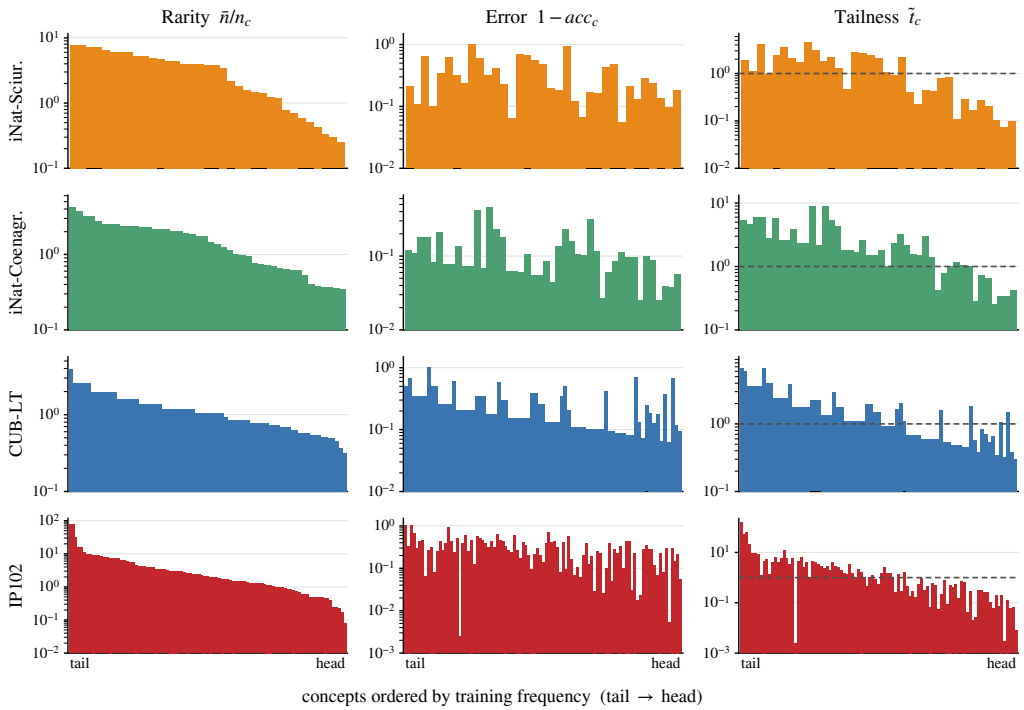


Figure 6: Rarity, error, and the calibrated tailness on the training set of each benchmark. Classes are sorted from the rarest to the most frequent, and all y-axes are in log scale. TARS amplifies advantages for classes above the dashed line at $\tilde{t}_c = 1$ and retains the vanilla GRPO advantage scale for the remaining classes.