Why Does RL Generalize Better Than SFT? A Data-Centric Perspective on VLM Post-Training

Supplementary Material

1. Implementation Details

For all experiments, we utilize the ms-swift [3] framework. The maximum pixel count is set to 802,816 (corresponding to an image size of 896 × 896) for all models. To adhere to this maximum pixel constraint, images are resized to ensure they do not exceed 896 × 896 pixels, and the labels for visual grounding are adjusted accordingly.

2. Comparison with Curriculum Learning.

Curriculum learning seeks to enhance model training by gradually introducing examples in a meaningful order, typically progressing from easier to more difficult instances [1]. Within the field of VLMs, curriculum learning has shown considerable promise for improving model performance and generalization [2]. While curriculum learning primarily focuses on the sequencing of training data, our study instead centers on enhancing generalization by filtering data.

3. Evaluation on Larger Training Set.

We further investigate the scalability of our approach under increased data conditions. To this end, we construct a larger subset of ImageNet-1K containing 200 classes with 500 samples each, yielding 100k training examples. ID evaluation is conducted on the corresponding 200-class test split. As shown in Table 1, DC-SFT (SFT-M) consistently outperforms standard SFT on OOD benchmarks and achieves performance on par with or surpassing GRPO across both model scales. These results reinforce the conclusions drawn from prior experiments and confirm that DC-SFT remains effective as training data increases.

Table 1. ID and OOD (gray background) performance (%) of different post-training paradigms with 100k training samples.

Model	Method	ImageNet	ImageNet-R	ImageNet-A
Qwen2.5-VL-3B	SFT	**93.03**	59.33	46.00
	GRPO	91.64 (-1.39)	65.79 (+6.46)	48.33 (+2.33)
	SFT-M	90.69 (-2.34)	**70.09** (+10.76)	**48.92** (+2.92)
Qwen2.5-VL-7B	SFT	**94.26**	63.16	52.44
	GRPO	93.16 (-1.10)	66.19 (+3.03)	53.86 (+1.42)
	SFT-M	92.16 (-2.10)	**66.94** (+3.78)	**54.20** (+1.76)

4. Data Details

4.1. Data Distribution

Table 2 presents the distribution of data across different difficulty levels for Qwen2.5-VL-3B and Qwen2.5-VL-7B. It can be observed that hard examples constitute only a small fraction of the overall data. Despite this, they have a substantial impact on the model's out-of-distribution performance, as demonstrated in the main text.

Table 2. Proportion of data of different difficulty levels.

Model	Dataset	Easy	Medium	Hard
Qwen2.5-VL-3B	ImageNet	43.4%	41.8%	14.8%
	RefCOCO	61.6%	34.6%	3.8%
Qwen2.5-VL-7B	ImageNet	49.0%	37.5%	13.5%
	RefCOCO	54.0%	40.3%	5.6%

4.2. Data Difficulty and Training Loss

Figure 1 shows the per-step training loss for Qwen2.5-VL-7B trained on easy, medium, and hard subsets from both ImageNet and RefCOCO. The results demonstrate that harder examples in our difficulty taxonomy consistently lead to higher training loss within SFT. This suggests that our difficulty taxonomy aligns well with a classification based on cross-entropy loss values.

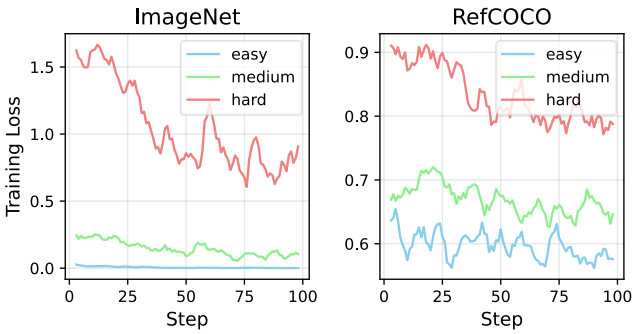

Figure 1. Training loss observed during SFT training on data subsets of varying difficulty.

4.3. Data Case

Figure 2 presents example images from each difficulty level, illustrating the characteristics of easy, medium, and hard instances. As can be seen, easy cases typically involve clear,

unambiguous visual cues (e.g., distinct objects or simple scenes). In contrast, hard cases often exhibit noise, ambiguity, or complex contextual relationships, such as identifying a specific object among visually similar items or interpreting subtle spatial descriptions.

Figure 2. Example images from different difficulty levels in the dataset. Hard examples tend to be more ambiguous, noisy, or require complex spatial reasoning.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 1

[2] Zihao Sheng, Zilin Huang, Yansong Qu, Yue Leng, Sruthi Bhavanam, and Sikai Chen. Curricuvlm: Towards safe autonomous driving via personalized safety-critical curriculum learning with vision-language models. *arXiv preprint arXiv:2502.15119*, 2025. 1

[3] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. 1