Learning What Matters: Prioritized Concept Learning via Relative Error-driven Sample Selection

Shivam Chandhok [*1 2 3] **Qian Yang** [*1 4] **Oscar Mañas** [1 4] **Kanishk Jain** [1 4] **Leonid Sigal** [2 3 5]
Aishwarya Agrawal [1 4 5]

Abstract

Instruction tuning has been central to the success of recent vision-language models (VLMs), but it remains expensive—requiring large scale datasets, high-quality annotations and large-compute budget. We propose **PR**ioritized c**O**ncept learnin**G** via **R**elative **E**rror-driven **S**ample **S**election – **PROGRESS** – a data-efficient framework that enables VLMs to dynamically select what to learn next based on their evolving needs during training. At each stage, the model tracks its learning progress across skills and selects the most *informative* samples: those it *has not already mastered* and *are not too difficult to learn* at the current state of training. This strategy effectively controls skill acquisition and the order in which skills are learned. Unlike prior works, PROGRESS requires no upfront answer annotations, querying answers only on a *need basis*, avoids reliance on additional supervision from auxiliary VLM, or compute-heavy gradient computations for data selection. Experiments across multiple instruction-tuning datasets of demonstrate that PROGRESS consistently outperforms state-of-the-art baselines with much less data and supervision.

1. Introduction

Multimodal vision-language models (VLMs) such as GPT-4V (OpenAI et al., 2024), Gemini (Team et al., 2023), LLaVA (Liu et al., 2023b;a), and InternVL (Chen et al., 2024b) demonstrate impressive general-purpose capabilities across tasks like image comprehension and visual question answering. Much of this success stems from large-scale fine-tuning on high-quality image-text corpora, particularly visual instruction-tuning (IT) datasets (Zhang et al., 2023; Xu et al., 2024), which significantly enhance instruction-following and reasoning abilities.

However, such pipelines are increasingly resource-intensive—annotation-heavy when relying on human-labeled supervision (*e.g.*, bounding boxes, object tags) and monetarily costly when generating instructions via proprietary models like GPT-4 (Liu et al., 2023b;a), alongside significant computational overhead. These factors make such pipelines increasingly inaccessible to individual researchers and smaller academic labs. More importantly, it is unclear whether the entirety of these large corpora is necessary for strong VLM performance.

To this end, we investigate how to select the *most informative* visual instruction-tuning (IT) samples based on the model's own evolving learning state. We ask: *Can VLMs indicate what they can most effectively learn at a give stage of training?* Inspired by curriculum learning, we develop a framework in which the model periodically self-evaluates its current knowledge and identifies the skills it is ready to acquire next—those that would most benefit its learning progress. Specifically, we track the relative change in skill performance across iterations to estimate where learning improves fastest, encouraging the model to prioritize these skills. We hypothesize that this enables the VLM to actively select training samples that are most informative: those that are *not already mastered* by the model, and are *not too difficult* for the model to learn at its current stage.

Experimental results across multiple instruction-tuning datasets of varying scale demonstrate that PROGRESS achieves up to **99–100%** of the full-data performance while using only **16–20%** of the labeled training data.

Our contributions are as follows:

- We propose **PROGRESS**, a dynamic, progress-driven framework for selecting the most informative samples during VLM instruction tuning—based on relative improvement across automatically discovered skills.
- Our method achieves near full-data performance using only 16–20% supervision across multiple

[*]Equal contribution [1]Mila - Québec AI Institute [2]University of British Columbia [3]Vector Institute for AI [4]Université de Montréal [5]Canada CIFAR AI Chair . Correspondence to: Shivam Chandhok <chshivam@cs.ubc.ca>, Qian Yang <qian.yang@mila.quebec>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

instruction-tuning datasets of varying scale—including the widely used LLaVA-665k dataset and Vision-Flan dataset—demonstrating strong data efficiency.

- It generalizes effectively across architectures, showing strong results on widely used - LLaVA-v1.5-7B, higher-capacity models like LLaVA-v1.5-13B and newer designs such as Qwen2-VL, while consistently outperforming competitive baselines and prior methods.

2. Related Work

Data Efficient Learning for VLMs. Efficient VLM training often relies on coreset selection using static metrics such as CLIP-Score (Hessel et al., 2022), or scoring function such as EL2N (Paul et al., 2021), perplexity (Marion et al., 2023), or entropy (Coleman et al., 2019), or via learned scoring networks (Chen et al., 2024a). These approaches typically perform selection only once before training and fail to adapt to evolving model needs. Moreover, static scores may miss key data modes, reducing diversity; harming generalization, as shown in prior work (Lee et al., 2024).

Other prominent work selects skill-diverse samples using reference VLMs—auxiliary models that themselves require large-scale instruction-tuning data to be effective. Methods like COINCIDE (Lee et al., 2024) select skill-diverse samples by clustering internal activations from an additional trained auxiliary VLM (e.g., TinyLLaVA (Zhou et al., 2024))—requiring full dataset annotations, an additional trained model, and manual intervention for activation selection—making them resource-intensive and less scalable.

Gradient-based selection (Wu et al., 2025; Liu et al., 2024b), while principled, demands high memory and compute (e.g., ICONS requires 100+ GPU hours), contradicting the goal of efficiency. Some also assume access to explicit knowledge of target task distribution in the form of sample(e.g ICONS (Wu et al., 2025)), which is rarely available in real-world VLM training.

3. Problem Setting and Overall Framework

Problem Setting. We now formally introduce our data-efficient learning setting for training VLMs. We denote an image by I, a question by Q, forming an image-question pair $(I, Q) \in \mathbb{U}$, where $\mathbb{U}$ is an unlabeled pool of such pairs. Unlike previous efficient learning methods, we do not assume access to the corresponding answers $A \in \mathbb{A}$ for all pairs in $\mathbb{U}$, and thus refer to this pool as unlabeled. The learner is provided with: (1) the unlabeled pool $\mathbb{U}$; and (2) a fixed answer budget b, specifying the maximum number of pairs from $\mathbb{U}$ for which it can query an answer $A \in \mathbb{A}$ and use for training, where $|\mathbb{A}| = b \ll |\mathbb{U}|$. The goal is to learn a vision-language model $\texttt{VLM}(A \mid I, Q)$ that can accurately predict an answer for a new image-question pair, while only using b selected and labeled samples during training. The central challenge lies in identifying the *most informative* (I, Q) pairs to annotate within the constrained budget b, such that the resulting model trained on these (I, Q, A) pairs performs comparably to one trained on the fully labeled dataset.

	Random	Perplexity	EL2N	Sem-DeDup	Self-Filter	ICONS*	COINCIDE	PROGRESS
1. Dynamic Selection	✗	✗	✗	✗	✗	✗	✗	✓
2. Order of Skills	✗	✗	✗	✗	✗	✗	✗	✓
3. Additional VLM Access	✗	✗	✓	✓	✓	✗	✓	✗
4. Target Task Access	✗	✗	✗	✗	✗	✓	✗	✗
5. Heavy-Gradient Overhead	✗	✗	✗	✗	✗	✓	✗	✗
6. Diversity	✓	✗	✗	✓	✗	✓	✓	✓
7. Answer Budget	20 %	20 %	100 %	100 %	100 %	100 %	100 %	20 %
8. Training Budget	20 %	20 %	20 %	20 %	20 %	20 %	20 %	20 %

Figure 1. **Comparison with Prior Efficient Learning Methods for VLMs.** **Green** denote **desirable** properties for efficient learning, while **Red** indicate **limitations**. PROGRESS *satisfies all key desirable criteria* while requiring only 20% data.

Overall Framework. Our overall framework is shown in Figure 2. We employ a two-stage pipeline:

(1) Multimodal Concept Categorization. Given an unlabeled data pool $\mathbb{U}$ containing image-question pairs $(I, Q) \in \mathbb{U}$, we first partition $\mathbb{U}$ into a distinct set of K skills, assigning each sample (I, Q) to a specific skill. This categorization enables tracking the model's progress on individual skills and supports a self-paced training strategy where the model's own learning signals determine which skills to prioritize next.

(2) Prioritized Concept Learning. During training, the model periodically self-evaluates its knowledge by comparing its current performance to prior state, identifying skills where performance improves fastest relative to prior state. Samples (I, Q) from these skills are then selected and answer annotations $A \in \mathbb{A}$ are queried only for these selected samples.

Our model adaptively selects diverse, informative samples aligned with evolving learning needs, enabling efficient training with minimal supervision while controlling both skill acquisition and learning order.

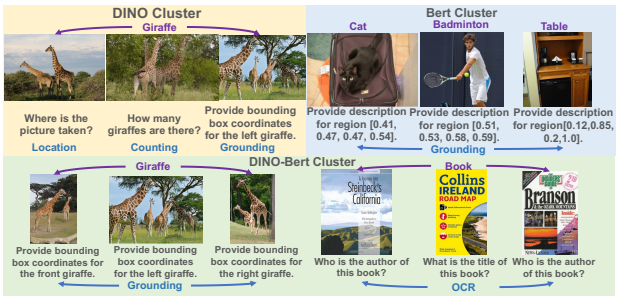

Figure 3. **Cluster Visualization.** Clustering using multimodal DINO-BERT features yields purer clusters with higher intra- and lower inter-cluster similarity than unimodal clustering

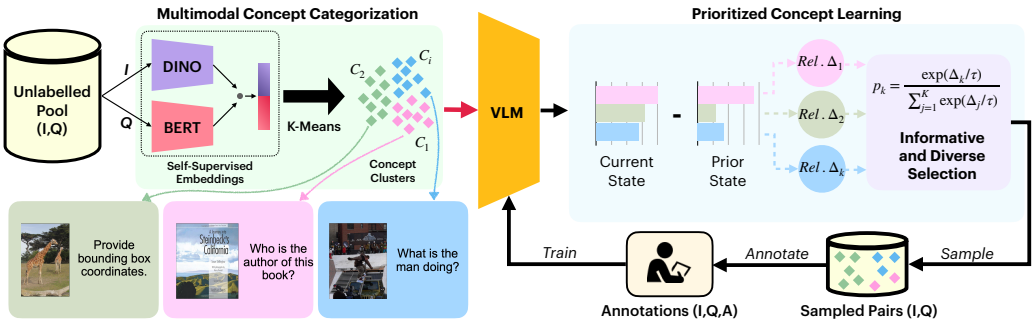

Figure 2. **Overall Pipeline.** The method consists of two stages: (1) *Multimodal Concept Categorization* clusters unlabeled image-question pairs into skill groups and (2) *Prioritized Concept Learning* dynamically selects samples from skills showing the highest learning progress.

3.1. Multimodal Concept Categorization

We begin by identifying diverse skills from the unlabeled data pool through a fully *unsupervised* concept categorization module that partitions $\mathbb{U}$ into K skill clusters using spherical k-means. Each sample $(I, Q) \in \mathbb{U}$ is assigned to a cluster based on cosine similarity from multimodal concatenated self-supervised DINO (Oquab et al., 2024) (for image I) and BERT (Devlin et al., 2019) (for text question Q) features. Jointly leveraging both modalities yields purer clusters with higher intra-cluster and lower inter-cluster similarity compared to unimodal partitioning (see Fig. 3)—enabling accurate tracking of skill-level progress during training. Unlike COINCIDE (closest best performing prior work) (Lee et al., 2024), our concept categorization framework is simpler, unsupervised, and more scalable. COINCIDE requires activations from fully trained additional VLM, ground-truth answers for full dataset, and human inspection to select appropriate activations for skill clustering. In contrast, our categorization is fully automated and practical, requires no labels, or manual introspection.

3.2. Prioritized Concept Learning: Can VLMs indicate what they can most effectively learn at a give stage of training?

Our goal is to guide the VLM to prioritize skills it can most readily learn and improve upon. Since human intuition about task difficulty may not align with model's difficulty in its feature and hypothesis space (Sachan & Xing, 2016), we adopt a self-paced strategy where the model's own learning progress determines what to learn next. Inspired by curriculum learning (Kumar et al., 2010; Sachan & Xing, 2016), we select the most informative samples—those that yield the greatest improvement in the model's objective (*e.g.*, **accuracy or loss**) relative to its prior state.

Formally, given an unlabeled pool $\mathbb{U} = \{(I, Q)\}$ partitioned into skill clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, we define the model's learning state at step t by its accuracy $\text{Acc}_k^{(t)}$ on each cluster k, computed over the training data seen by the model so far. The *relative change in performance* across steps quantifies learning progress per skill, which is then used to guide sample selection. We compute the expected accuracy improvement for each skill cluster between step t and $t - \gamma$:

$$\Delta_k = \frac{\text{Acc}_k^{(t)} - \text{Acc}_k^{(t-\gamma)}}{\text{Acc}_k^{(t-\gamma)} + \epsilon} \tag{1}$$

where ϵ ensures numerical stability. The score Δ_k captures how rapidly the model is improving on skill cluster k, serving as a proxy for sample *informativeness*. By prioritizing high Δ_k clusters , the model focuses on skills it can improve on most rapidly—*thereby enforcing a self-paced curriculum* that dynamically adapts to the model's learning state (Sachan & Xing, 2016)—controlling both the acquisition of skills and the order in which they are learned.

Only selected samples are annotated, forming the labeled set (I, Q, A) for training. This *need-based annotation* strategy avoids the costly requirement of full supervision used in prior coreset methods (such as COINCIDE (Lee et al., 2024)), offering a more scalable and efficient training.

However, naively selecting samples from only the highest-improvement cluster can hurt diversity by concentrating on a narrow skill set and leading to mode collapse—an issue known to degrade performance in prior work (Lee et al., 2024). To mitigate this, we propose to sample from multiple high Δ_k clusters in proportion to their relative improvement using a *temperature-controlled softmax*:

$$p_k = \frac{\exp(\Delta_k/\tau)}{\sum_{j=1}^{K} \exp(\Delta_j/\tau)} \tag{2}$$

Here, p_k is the probability of sampling from cluster k, and τ controls the sharpness of the distribution. Lower τ emphasizes top clusters but risks mode collapse by repeatedly sampling from a narrow skill set (higher informativeness, lower diversity); higher τ promotes broader sampling and better skill coverage. This balance between **informativeness and diversity** is critical for effective and robust learning (see ablation Fig. 4). The sampling budget at given step t is

Table 1. Comparison of coreset selection techniques with LLaVA-v1.5-7b on the LLaVA-1.5 dataset using 20% sampling ratio. Methods highlighted in orange require additional reference VLMs and ground-truth labels for coreset selection, while methods highlighted in light green do not require either. The benchmark results are highlighted with best and second best comparing models within respective categories with and without additional information.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Wild	SEED	AI2D	ChartQA	CMMMU	Rel. (%)
0 Full-Finetune	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	67.0	56.4	16.4	22.1	100
1 Self-Sup	74.9	59.5	46.0	67.8	49.3	83.5	1335.9	61.4	53.8	63.3	62.5	52.9	16.1	23.4	94.6
1 Self-Filter	73.7	58.3	53.2	61.4	52.9	83.8	1306.2	48.8	45.3	64.9	60.5	48.7	14.1	19.8	90.1
3 EL2N	76.2	58.7	43.7	65.5	53.0	84.3	1439.5	53.2	47.4	64.9	61.8	49.3	16.5	23.9	93.4
4 SemDeDup	74.2	54.5	46.9	65.8	55.5	84.7	1376.9	52.2	48.5	70.0	60.9	53.5	15.8	24.2	94.1
5 D2-Pruning	73.0	58.4	41.9	69.3	51.8	85.7	1391.2	65.7	57.6	63.9	62.1	52.5	15.3	22.3	94.8
6 COINCIDE	76.5	59.8	46.8	69.2	55.6	86.1	1495.6	63.1	54.5	67.3	62.3	53.3	16.1	24.3	97.8
7 Random	75.7	58.9	44.3	68.5	55.3	84.7	1483.0	62.2	54.8	65.0	61.7	50.2	15.1	21.9	95.0
8 CLIP-Score	73.4	51.4	43.0	65.0	54.7	85.3	1331.6	55.2	52.0	66.2	61.0	49.1	14.3	20.3	90.6
9 Perplexity	75.8	57.0	47.8	65.1	52.8	82.6	1341.4	52.0	45.8	68.3	60.8	48.7	14.5	20.9	91.1
PROGRESS															
10 Loss as Obj.	75.7	58.6	49.6	70.1	55.1	86.3	1498.4	62.5	55.5	65.5	63.4	53.3	17.3	23.7	98.4
11 Accuracy as Obj.	75.2	58.8	53.4	69.9	55.1	85.9	1483.2	61.1	54.4	65.5	63.0	52.8	17.3	24.6	**98.8**

then allocated proportionally to p_k, and only the selected samples are annotated as (I, Q, A) triplets for training.

4. Experiments and Results

Experimental Setup *-See Appendix for more details*

1. PROGRESS is more effective than existing SOTA in data efficient learning.

Table 1 (Row 0-11) compares PROGRESS against state-of-the-art baselines for training LLaVA-v1.5-7B on LLaVA-665K dataset under a 20% data budget, following standard settings. PROGRESS achieves the highest relative performance (98.8%), outperforming all baselines, including those requiring access to ground-truth answers for the entire dataset and additional reference VLMs. In contrast, PROGRESS uses supervision only on a *need basis* for 20% of samples and relies solely on self-supervised features, yet reaching near-parity with full finetuning. PROGRESS also ranks among the top two methods on 8 out of 14 benchmarks, showing strong generalization across diverse tasks—*e.g.*, including perception-focussed VQAv2 (75.2), scientific questions and diagrams (ChartQA:17.3, AI2D:52.8), and object hallucination POPE (85.9). Notably, it exceeds full-data performance on VizWiz (53.4 vs. 47.8), SQA-I (69.9 vs. 68.4), MME (1483.2 vs. 1476.9), ChartQA (17.3 vs. 16.4) and CMMMU (24.6 vs. 22.1). These results demonstrate effectiveness the PROGRESS as a dynamic and fully automated alternative for efficient VLM training under limited supervision.

2. PROGRESS generalizes across datasets (Vision-Flan) and across architectures (Qwen-2-VL and LLaVA-1.5-13B) *-See Appendix for more details*

3. Balancing Diversity and Informativeness

To assess the importance of balancing informativeness and diversity (Eqn 2), we conduct an ablation study varying the *temperature parameter* τ in the softmax used for skill selection (Eqn 2). As shown in Figure 4, a *high temperature of 1.0* yields the **best overall performance** (98.8% relative score), striking an effective balance between prioritizing high-improvement clusters and maintaining skill diversity.

As the temperature decreases (*i.e.*, $\tau = 0.7, 0.5, 0.3$), performance consistently degrades, with the lowest temperature yielding only 92.8%—a significant drop of over 6% in relative score. This decline confirms that *overly sharp sampling distributions* (low τ) lead to *mode collapse*, where the model repeatedly focuses on a narrow set of concepts and fails to generalize broadly.

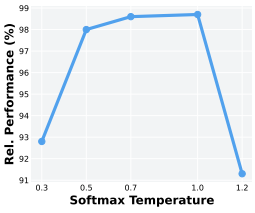

Figure 4. **Ablation of Softmax Temperature.** Both very-low and very-high temperatures lead to a significant performance drop.

Thus, we see that enforcing diversity helps improve performance. However, excessive diversity (τ = 1.2) is also not good as, in that case, the high-improvement clusters start losing their clear priority over other clusters.

5. Conclusion and Limitations

We propose PROGRESS, a dynamic and data-efficient framework for instruction-tuning VLMs that prioritizes learning based on progress across unsupervised skill clusters. By selecting samples that are both learnable and timely, it effectively controls skill acquisition and learning order. PROGRESS achieves near full-data performance using only 16–20% supervision, requires no auxiliary VLMs, and generalizes across architectures and datasets. While effectively orders and prioritizes more informative skills, it randomly samples within selected skills without finer-grained ranking, and the accuracy-based variant adds inference overhead (mitigated by a loss-based alternative), PROGRESS consistently outperforms prior methods with far less supervision and greater scalability.

References

Abbas, A. K. M., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Chen, R., Wu, Y., Chen, L., Liu, G., He, Q., Xiong, T., Liu, C., Guo, J., and Huang, H. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4156–4172, 2024a.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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*, pp. 24185–24198, 2024b.

Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., and Zaharia, M. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Ge, Z., Xinrun, D., Bei, C., Yiming, L., Tongxu, L., Tianyu, Z., Kang, Z., Yuyang, C., Chunpu, X., Shuyue, G., Haoran, Z., Xingwei, Q., Junjie, W., Ruibin, Y., Yizhi, L., Zekun, W., Yudong, L., Yu-Hsuan, T., Fengji, Z., Chenghua, L., Wenhao, H., and Jie, F. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *GitHub repository*, 2024.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Kumar, M., Packer, B., and Koller, D. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

Lee, J., Li, B., and Hwang, S. J. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5060–5080, 2024.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://openreview.net/forum?id=xozJw0kZXF.

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., and Liu, K. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Liu, Z., Zhou, K., Zhao, W. X., Gao, D., Li, Y., and Wen, J.-R. Less is more: High-value data selection for visual instruction tuning, 2024b. URL https://arxiv.org/abs/2403.09559.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Maharana, A., Yadav, P., and Bansal, M. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Masry, A., Do, X. L., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

OpenAI, Achiam, J., and et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Sachan, M. and Xing, E. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 453–463, 2016.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wu, X., Xia, M., Shao, R., Deng, Z., Koh, P. W., and Russakovsky, O. Icons: Influence consensus for vision-language data selection, 2025. URL https://arxiv.org/abs/2501.00654.

Xu, Z., Feng, C., Shao, R., Ashby, T., Shen, Y., Jin, D., Cheng, Y., Wang, Q., and Huang, L. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*, 2024.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., and Huang, L. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

6. Appendix

6.1. Experimental Setup

Datasets and Models. To demonstrate effectiveness and generalizability of our approach across different scales of instruction-tuning (IT) data, we conduct experiments on two IT datasets: Visual-Flan-191K (Xu et al., 2024) and the larger-scale LLaVA-665K (Liu et al., 2023b) containing ~ 0.6 million samples. For the target VLMs, we use LLaVA-v1.5-7B (Liu et al., 2023b), following prior works. Additionally, we use LLaVA-v1.5-13B (Liu et al., 2023b) to test scalability and Qwen2-VL-7B (Wang et al., 2024) to test architecture generalization towards newer VLMs.

Implementation Details. Following standard protocol(Lee et al., 2024), we adopt LoRA (Hu et al., 2021) for training using the official hyperparameters from LLaVA-1.5. For the accuracy-based variant, we estimate cluster-wise accuracy using an LLM judge that compares the VLM output to ground-truth answers for samples in given cluster—though this is not required for our loss-based variant. Our setup follows standard evaluation protocols from prior work, ensuring consistency and fair comparison.

Baselines. We compare PROGRESS against strong baselines spanning five major categories: (1) scoring function based methods (CLIP-Score, EL2N(Paul et al., 2021), Perplexity (Marion et al., 2023)); (2) deduplication-based selection (SemDeDup (Abbas et al., 2023)); (3) graph-based methods (D2-Pruning (Maharana et al., 2024)); (4) learned static selectors (Self-Filter (Chen et al., 2024a)); and (5) concept-diversity approaches (COINCIDE (Lee et al., 2024), Self-Sup (Sorscher et al., 2022)). We also include *Random*—a simple yet competitive baseline shown to perform well due to its diversity—and *Full-Finetune*, representing the performance upper bound with full supervision.

Evaluation Benchmark. We evaluate our approach on a diverse suite of 14 vision-language benchmarks targeting different skills: perceptual reasoning (VQAv2 (Goyal et al., 2017), VizWiz (Gurari et al., 2018)), textual reasoning (TextVQA (Singh et al., 2019)), compositional reasoning (GQA (Hudson & Manning, 2019)), object hallucinations (POPE (Li et al., 2023b)), multilingual understanding (MMBench-cn (Liu et al., 2024a), CMMMU (Ge et al., 2024)), instruction-following (LLaVA-Bench(Liu et al., 2023b)), fine-grained skills (MME (Liang et al., 2024), MMBench-en (Liu et al., 2024a), SEED (Li et al., 2023a)), and scientific questions and diagrams (SQA-I (Lu et al., 2022), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022)).

Evaluation Metrics. We use standard evaluation metrics used by previous work to ensure consistency. Specifically,

[2]Reproduced with official code.

Table 2. **Scalability to Larger Models** - Comparison of coreset selection techniques for training larger LLaVA-v1.5-13B on the LLaVA-665K dataset using 20% sampling ratio.[2] Methods highlighted in orange require additional reference VLMs and 100% dataset annotations for coreset selection, while methods highlighted in light green do not require either. The benchmark results are highlighted with best and second best models within the respective categories (i.e, with and without utilizing additional information). The best and the second best relative score are in **bold** and underlined, respectively.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Wild	SEED	AI2D	ChartQA	CMMMU	Rel. (%)
LLaVA-v1.5-13B															
1 Full-Finetune	80.0	63.3	58.9	71.2	60.2	86.7	1541.7	68.5	61.5	69.5	68.3	60.1	19.3	22.1	100
2 Self-Sup	76.3	60.5	50.0	70.2	52.7	85.4	1463.8	63.7	57.6	64.9	65.2	53.3	17.2	23.2	93.8
3 Self-Filter	75.0	59.8	48.6	69.5	55.8	84.5	1446.9	58.8	51.8	69.1	65.3	52.4	16.9	23.1	92.6
4 EL2N	77.2	59.6	54.8	69.9	56.1	84.1	1531.0	59.3	52.3	65.8	65.7	53.9	17.0	24.4	94.4
5 SemDeDup	75.6	57.5	48.3	70.5	57.7	85.3	1397.6	59.0	51.1	68.7	64.9	53.2	16.8	24.6	92.9
6 D2-Pruning	73.9	60.5	49.8	70.4	55.2	84.9	1463.0	67.3	59.9	66.5	65.9	53.4	16.9	23.5	94.7
7 COINCIDE	77.3	59.6	49.6	69.2	58.0	87.1	1533.5	64.5	56.6	66.4	65.9	52.9	18.4	25.0	95.9
8 Random	76.7	60.5	48.0	68.8	57.7	84.8	1484.9	62.8	55.2	68.6	65.5	57.9	17.1	24.3	95.0
9 CLIP-Score	75.3	52.6	42.2	69.7	57.3	85.4	1426.3	60.4	54.0	68.1	63.3	52.8	17.4	23.7	91.8
10 Perplexity	77.0	58.5	48.2	68.7	54.8	83.1	1508.8	57.5	50.3	68.7	64.7	53.1	17.6	23.8	92.7
PROGRESS															
11 Loss as Obj.	76.8	59.7	54.6	70.4	58.0	87.2	1458.3	63.8	56.9	69.9	65.1	58.0	17.9	24.6	**96.8**
12 Accuracy as Obj.	76.9	58.9	53.0	70.1	57.5	87.1	1497.6	63.9	57.6	67.3	65.4	57.7	18.0	24.5	96.5

we use **average relative performance** (Lee et al., 2024)to provide a unified measure for generalization. For each benchmark, relative performance is defined as: Rel. $= \left(\frac{\text{Model Performance}}{\text{Full Finetuned Performance}}\right) \times 100\%$ This normalization allows us to normalize for the differences in the difficulty levels of different benchmarks following previous work.

6.2. Results and Analysis

Scalability to Larger Models. To assess scalability, we use PROGRESS to train the larger LLaVA-v1.5-13B model under the same 20% data budget, testing whether our method developed for LLaVA-v1.5-7B transfers effectively to a higher-capacity model without hyperparameter tuning. As shown in Table 2, PROGRESS achieves a relative performance of 96.8%, outperforming all baselines. Beyond aggregate gains, PROGRESS ranks among the top-2 methods on 8 out of 14 benchmarks compared with all baselines, demonstrating strong generalization.

Architectures and Dataset Generalization. In Table 3, we test generalization of PROGRESS across different VLM architecture and IT dataset with accuracy as signal. For **architecture generalization**, we use newer Qwen2-VL-7B and train it on the LLaVA-665K dataset using the same 20% data budget and identical hyperparameters. We compare PROGRESS with two of the strongest (highest performing) established baselines—Random Sampling and COINCIDE—across multiple multimodal benchmarks. PROGRESS achieves the highest overall relative performance of 100% and ranks first or second on 9 out of 11 benchmarks (Tab. 3, **top**). For **dataset generalization**, we report LLaVA-v1.5-7B on Vision-Flan dataset under a stricter 16.7% annotation budget to assess generalization in low-resource settings. PROGRESS achieves the highest overall relative performance of 99.0%, outperforming COINCIDE (95.8%) and Random (95.0%) and ranks first or second on 8 out of 11 benchmarks (Tab. 3, **bottom**). The scalability and generalization are achieved without requiring any model-specific tuning. These results underscore the calability and generalization of PROGRESS , making it a practical solution for efficient training across diverse architecture and datasets.

How effective is PROGRESS under different sampling ratios? We show relative performance on the Vision-FLAN dataset under different sampling ratios (even lower than 16.7 % considered in Tab. 3) in Figure 5. PROGRESS consistently outperforms strongest baselines - Random and COINCIDE across different sampling ratios, highlighting its effectiveness.

6.3. Investigating the effectiveness of different components of PROGRESS

In this section, we provide further insights into the learning behaviour of PROGRESS . All experiments use LLaVA-v1.5-7B and LLaVA-665K dataset with 20% sampling ratio and accuracy as the objective unless otherwise specified.

How effective is our Selection Policy? We evaluate the efficacy of PROGRESS (sampling based on relative accuracy

Table 3. **Architecture and Dataset Generalization.** For Architecture Generalization, we report Qwen2-VL-7B on the LLaVA-665K dataset using 20% sampling ratio. For Dataset Generalization, we report LLaVA-v1.5-7B on Vision-Flan dataset using 16.7% sampling ratio following prior work.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Wild	SEED	Rel. (%)
Architecture Generalization (Qwen2-VL-7B)												
Full-Finetune	77.4	61.7	45.5	81.4	59.7	84.3	1567.9	76.1	75.1	84.8	66.9	100
Random	76.2	60.1	43.6	81.4	58.7	83.7	1556.8	76.8	74.5	81.7	67.6	98.7
COINCIDE	76.7	60.2	45.4	81.7	59.4	83.6	1583.5	77.4	76.2	80.5	67.9	99.6
PROGRESS	76.2	60.5	47.1	82.3	58.0	84.3	1560.1	77.2	72.9	87.1	67.6	**100.0**
Dataset Generalization (Vision-Flan-191K)												
Full-Finetune	69.4	46.0	49.7	59.9	34.1	85.1	1306.1	49.1	51.7	35.7	53.3	100
Random	66.0	43.8	52.2	62.1	39.7	82.7	1072.2	48.7	43.7	40.4	28.7	95.0
COINCIDE	66.3	43.6	51.0	63.8	35.2	81.9	1222.2	56.7	45.5	31.1	37.5	95.8
PROGRESS	65.5	44.0	53.6	62.5	42.0	82.9	1040.9	43.6	47.4	43.2	45.3	**99.0**

Table 4. **Ablation of Selection Policy.** Performance comparison of different selection policies with the same warm-up strategy.

	Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Wild	SEED	AI2D	ChartQA	CMMMU	Rel. (%)
0	Full-Finetune	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	67.0	56.4	16.4	22.1	100
1	Warm-up Only	73.1	55.9	43.8	67.9	54.2	85.4	1410.3	58.5	52.7	64.6	60.5	52.4	16.1	24.5	94.6
2	Random	75.7	59.0	43.8	68.8	54.9	85.6	1414.2	61.9	54.9	66.2	63.3	48.6	17.3	25.2	96.8
3	Easiest	72.0	54.8	50.2	67.1	51.6	85.7	1407.4	57.0	52.6	65.2	59.5	50.1	12.3	22.8	92.3
4	Medium	69.3	52.5	46.0	68.3	50.8	85.4	1307.6	54.6	48.7	62.5	57.7	47.6	14.3	26.1	91.1
5	Hardest	72.8	54.8	52.1	61.3	50.5	85.4	1364.8	37.9	34.5	67.5	54.1	41.4	15.8	25.9	88.5
PROGRESS																
6	Loss as Obj.	75.7	58.6	49.6	70.1	55.1	86.3	1498.4	62.5	55.5	65.5	63.4	53.3	17.3	23.7	98.4
7	Accuracy as Obj.	75.2	58.8	53.4	69.9	55.1	85.9	1483.2	61.1	54.4	65.5	63.0	52.8	17.3	24.6	**98.8**

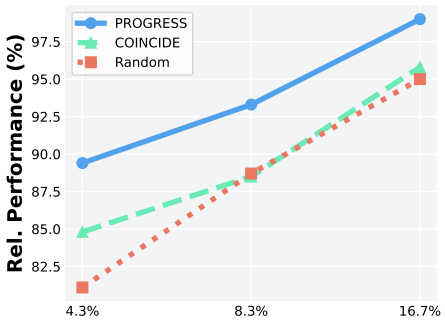

Figure 5. **Ablation of Sampling Ratio**. Relative performance on Vision-Flan dataset under different sampling ratio.

change - Sec 3.2) by comparing it against other competitive selection strategies: **Random Sampling**, **Easiest-Sampling** (selecting clusters with highest absolute accuracy at given time step), **Medium-Sampling** (selecting moderate accuracy clusters), and **Hardest-Sampling** (selecting lowest accuracy clusters). As shown in Tab. 4, PROGRESS achieves the highest relative score (98.8%), ranking first on 8 out of 14 benchmarks and second on 4 others.

How important is the order of skill acquisition? Here we randomly shuffle PROGRESS-selected samples and perform training —thereby ablating importance of learning order. We find a 4.1% drop in relative performance, underscoring the importance of introducing appropriate skills at the right time.