

Filter Images First, Generate Instructions Later: Pre-Instruction Data Selection for Visual Instruction Tuning

Bardia Safaei^{1*}, Faizan Siddiqui², Jiacong Xu¹, Vishal M. Patel¹, Shao-Yuan Lo²

¹Johns Hopkins University, ²Honda Research Institute USA

{bsafaei1, jxu155, vpatel136}@jhu.edu {faizan.siddiqui, shao-yuan.lo}@honda-ri.com

A1. Limitations and Potential Societal Impacts

Limitations. Currently, our approach utilizes the reference instructions for task-importance estimation. However, given the multimodal nature of LVLMS, these instructions may also be leveraged to improve the selection process on unlabeled images. We leave this limitation as an interesting direction for future work.

Potential Negative Societal Impacts. The proposed method could potentially enable malicious actors to fine-tune open-source LVLMS more easily for illegal purposes. The integration of machine learning security mechanisms could be studied to mitigate such risks.

A2. Different Model Sizes and Architectures

PreSel adapts to various model sizes and architectures. In addition to the results presented in Table 1 and Table 2, which are based on the LLaVA-7B model using Vicuna-7B [2] as the LLM, we conduct two additional sets of experiments on the LLaVA-1.5 dataset, as shown in Table 7. In these experiments, we change the LLM to Vicuna-13B [2] and Llama-8B [3] to evaluate the transferability of the samples selected by PreSel across different model sizes and architectures. It is important to note that we directly use the samples selected by PreSel with LLaVA-7B as the reference model to fine-tune the additional LVLMS, without any further processing. From the table, it is evident that our selected samples are highly beneficial for fine-tuning LVLMS with different architectures or sizes. Specifically, PreSel outperforms Random by large margins of 1.7% and 1.2% in average relative performance for the LLaVA-Vicuna-13B and LLaVA-Llama-8B models, respectively.

A3. More Analysis on the Size of Reference Set

We leverage a small, randomly selected set of image-instruction pairs as the reference set, \mathcal{D}_{ref} , to estimate task-importance values during the Task-Importance Estimation stage. By default, the size of \mathcal{D}_{ref} is set to 5% of the total size

of the VIT dataset, \mathcal{D} . In Table 8, we reduce the reference set size to just 1% of the total VIT dataset to evaluate the effect of $|\mathcal{D}_{\text{ref}}|$ on the average relative performance. The rest of the experimental settings are identical to those in Table 1 of the main paper: we set the sampling ratio to 15% and conduct experiments on the LLaVA-1.5 dataset. Our experiments show that the estimated task-importance values remain almost identical for both the 1% and 5% cases. Consequently, the final performance is robust to the size of the reference set, as shown in Table 8.

A4. Baselines

In this section, we provide details about the baselines with which we compared our method.

- **CLIP-Score** [9]. In CLIP-Score, the cosine similarity between the images and their corresponding textual instructions is used for selection. In our experiments, we select samples with high similarity.
- **TypiClust** [4]. TypiClust is an active learning approach originally designed for multi-round data selection for image classification under low-budget regimes. It adaptively clusters the unlabeled data and selects the most typical samples from the clusters that have the highest number of unlabeled samples and the fewest already labeled samples. We adapt a single-round version of this method to our setting.
- **EL2N** [8]. This method uses the Error L2-Norm (EL2N) to quantify a sample’s informativeness. Specifically, it calculates the average L2-Norm distance between generated tokens and ground-truth text tokens to produce this score.
- **Perplexity** [7]. This method uses the average next-token prediction loss as a measure of the LVLMS’s uncertainty with respect to a sample. Following prior research [7], we select samples with medium score values instead of top values, as it leads to better results.
- **IFD** [6]. This method proposes the Instruction-Following Difficulty (IFD) metric to select samples with corresponding instructions that have a minimal impact on the model’s loss.

| Model | Method | VQAv2 | SQA-I | TextVQA | MME | MMBench | | SEED-Bench | MM-Vet | POPE | Rel. (%) |
|----------------|----------------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | | | | | en | cn | | | | |
| Vicuna-13B [2] | Full Finetune | 80.0 | 72.0 | 58.3 | 1400.4 | 68.2 | 60.7 | 68.0 | 34.3 | 87.1 | 100 |
| | Random | 76.2 | 69.9 | 56.8 | 1452.3 | 62.8 | 56.0 | 65.0 | 33.0 | 86.6 | 96.6 |
| | PreSel (Ours) | 76 | 70.8 | 57.0 | 1404.3 | 66.4 | 60.9 | 65.0 | 34.4 | 87.2 | 98.3 |
| Llama-3-8B [3] | Full Finetune | 81.4 | 78.2 | 63.8 | 1561.2 | 75.2 | 73.1 | 71.8 | 36.9 | 85.6 | 100 |
| | Random | 77.5 | 78.3 | 59.8 | 1455.0 | 72.5 | 69.1 | 68.3 | 35.7 | 84.9 | 96.0 |
| | PreSel (Ours) | 77.0 | 79.0 | 59.2 | 1501.4 | 73.3 | 70.6 | 68.5 | 37.7 | 84.6 | 97.2 |

Table 7. **Results for LLaVA-Vicuna-13B and LLaVA-Llama-8B models.** We report the performance of PreSel across different model architectures and sizes. The experiments are conducted on the LLaVA-1.5 dataset and the sampling ratio is set to 15% for both Random and PreSel methods. We directly use the samples acquired for the LLaVA-Vicuna-7B model to fine-tune the additional LVLMS in this experiment. The average relative performance across all evaluation benchmarks is reported.

| $\frac{ D_{\text{ref}} }{ D }$ | Method | VQAv2 | SQA-I | TextVQA | MME | MMB-E | MMB-C | SEED-Bench | MM-Vet | POPE | Rel. (%) |
|--------------------------------|---------------|-------|-------|---------|--------|-------|-------|------------|--------|------|----------|
| - | Full Finetune | 79.1 | 68.4 | 57.9 | 1417.6 | 66.0 | 58.9 | 66.8 | 30.0 | 87.5 | 100 |
| 5% | PreSel | 75.0 | 70.1 | 55.2 | 1457.7 | 64.8 | 56.5 | 63.8 | 29.6 | 85.4 | 97.9 |
| 1% | PreSel | 74.9 | 69.7 | 54.2 | 1436.5 | 63.1 | 57.4 | 63.9 | 30.0 | 86.5 | 97.7 |

Table 8. The effect of $|D_{\text{ref}}|$ on average relative performance. The experiments are conducted on the LLaVA-1.5 dataset and the sampling ratio is set to 15%. “Full Finetune” denotes a LLaVA-1.5-7B model fine-tuned on the entire LLaVA-1.5 image-instruction pairs.

- **Self-Filter [1].** The Self-Filter method involves initially training a score-net while fine-tuning LVLMS on the entire VIT dataset. Afterwards, the score-net is employed to choose a subset of data for a subsequent VIT round. Nonetheless, this two-step procedure escalates the total training expense, opposing the purpose of data selection.
- **COINCIDE [5].** This method first utilizes the TinyLLaVA-2B [10] model to extract both image and instruction features from multiple layers. It then concatenates all features, performs spherical clustering on them, and selects samples from clusters proportional to their overall transferability.

A5. ShareGPT Data in LLaVA-1.5 Dataset

The LLaVA-1.5 dataset includes a task named ShareGPT, comprised of text-only instructions generated by the ShareGPT model. In our experiments, we have excluded these text-only instructions as our primary focus is to select the most beneficial ‘images’ for visual instruction tuning. One can easily choose a subset of ShareGPT data or use all of it along with our selected VIT samples for fine-tuning LVLMS.

A6. Detailed Algorithm for PreSel

We elaborate on the details of our PreSel pre-instruction data selection approach in Algorithm 1.

References

- [1] Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 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*, 2024. 2
- [2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. <https://lmsys.org/blog/2023-03-30-vicuna>. 1, 2
- [3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 2
- [4] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning*, 2022. 1
- [5] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2024. 2
- [6] Ming Li, Yong Zhang, Zhitao Li, Jiu-hai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Conference of*

Algorithm 1 Our Proposed PreSel Approach

- 1: **Input:**
 - 2: unlabeled images $\mathcal{D} = \bigcup_{i=1}^M T_i$ where T_i is the i -th task, number of vision tasks M , a lightweight vision encoder (DINOv2), our LVLM, the small randomly selected reference set \mathcal{D}_{ref}
 - 3: **Process:**
 - 4: $\mathcal{D}_S \leftarrow \emptyset$ # Initialization for the Selected Subset
 - 5: # Task-Importance Estimation
 - 6: for each $(I, Q, R) \in \mathcal{D}_{ref}$: calculate $\mathcal{L}_{R|Q,I}$ and $\mathcal{L}_{R|I}$ via Eq. 1 and Eq. 2 to obtain IRS via Eq. 3
 - 7: **for** $i = 1, \dots, M$ **do**
 - 8: calculate $s(T_i)$ and $w(T_i)$ via Eq. 4 and Eq. 5
 - 9: # Task-wise Cluster-based Selection
 - 10: **for** $i = 1, \dots, M$ **do**
 - 11: $\mathcal{D}_{T_i} \leftarrow \emptyset$ # Selected Images for Task T_i
 - 12: for unlabeled images in T_i , extract visual features using DINOv2 encoder
 - 13: cluster visual features in T_i into C clusters $\{A_c^i\}_{c=1}^C$ where $C = \frac{|T_i|}{100}$
 - 14: # Intra-Cluster Selection
 - 15: **for** $c = 1, \dots, C$ **do**
 - 16: for A_c^i , set the cluster budget n_c via Eq. 6
 - 17: calculate s_{n_c} for all images within A_c^i via Eq. 7
 - 18: rank images in A_c^i based on s_{n_c} and select the top n_c samples with the highest values $\mathcal{D}_{T_i}^c$
 - 19: $\mathcal{D}_{T_i} \leftarrow \mathcal{D}_{T_i} \cup \mathcal{D}_{T_i}^c$
 - 20: acquire instructions for images in \mathcal{D}_{T_i}
 - 21: $\mathcal{D}_S \leftarrow \mathcal{D}_{T_1} \cup \mathcal{D}_{T_2}, \dots, \cup \mathcal{D}_{T_M} \cup \mathcal{D}_{ref}$
 - 22: # Visual Instruction Tuning
 - 23: fine-tune the LVLM
 - 24: **Return** the fine-tuned LVLM
-

the Nations of the Americas Chapter of the Association for Computational Linguistics, 2024. 1

- [7] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023. 1
- [8] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Conference on Neural Information Processing systems*, 2021. 1
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1
- [10] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 2