Data Selection for Vision-Language Models

Paper: Concept-skill Transferability-based Data Selection for Large Vision-Language Models (*EMNLP* 2024).

Motivation

- Finetuning on full VIT datasets is prohibitively expensive for many users'

- Different VL tasks share overlapping concept-skill compositions

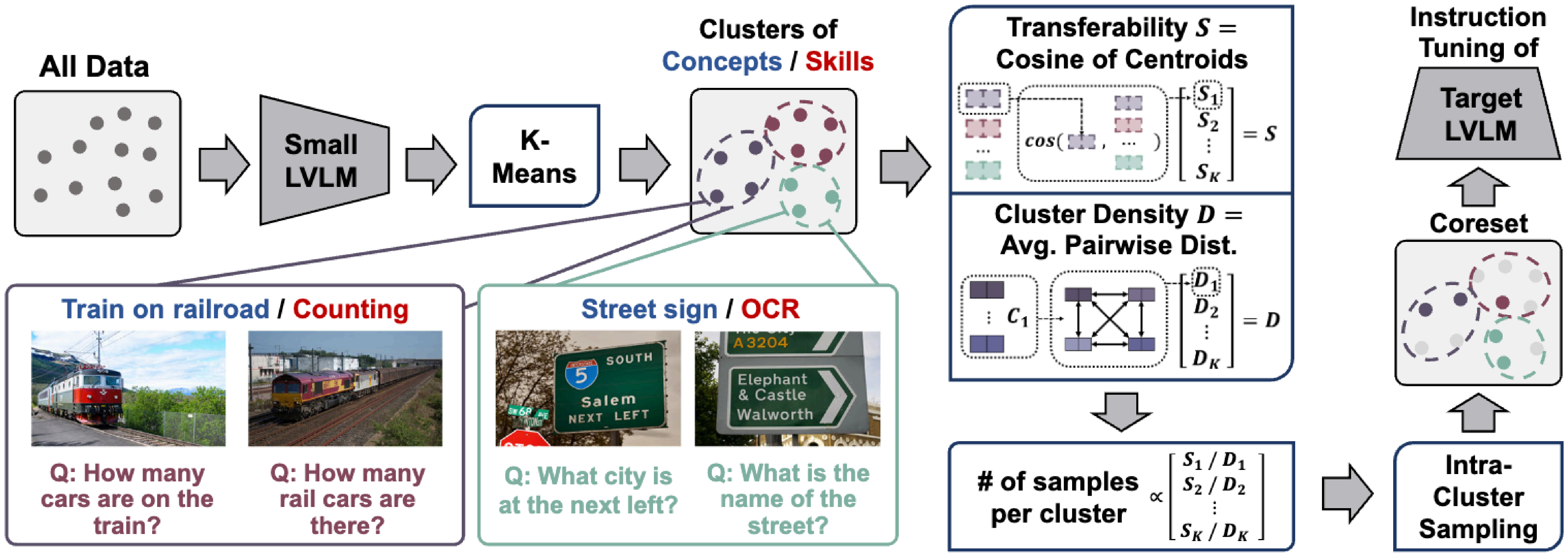
- Single-metric selection (e.g., EL2N, Self-Filter) yields biased coresets across tasks with different score distributions, harming diversity

RQ1: How to select a small, diverse, and transferable *coreset* for visual instruction tuning (VIT) of LVLMs to preserve generalization while reducing compute?

Lee, J., Li, B., & Hwang, S. J. (2024). Concept-skill Transferability-based Data Selection for Large Vision-Language Models. Conference on Empirical Methods in Natural Language Processing, 5060–5080.

RQ1: How to select a small, diverse, and transferable *coreset* for visual instruction tuning (VIT) of LVLMs to preserve generalization while reducing compute?

Qualitative analysis indicates the clusters coincide with concept-skill compositions.



Lee, J., Li, B., & Hwang, S. J. (2024). Concept-skill Transferability-based Data Selection for Large Vision-Language Models. Conference on Empirical Methods in Natural Language Processing, 5060–5080.

RQ1: How to select a small, diverse, and transferable *coreset* for visual instruction tuning (VIT) of LVLMs to preserve generalization while reducing compute?

Concept-skill compositions: Combinations of visual-semantic concepts and associated skills (tasks or instructions) that represent meaningful clusters of vision-language data relevant for training.

Hypotheses:

- Neuron activations from a small VLM can cluster VIT data into meaningful concept-skill compositions.
- Allocating more samples to high-transferability and low-density clusters improves efficiency without sacrificing diversity.
- Cluster transferability positively correlates with centroid cosine similarity, enabling a cheap proxy.

Lee, J., Li, B., & Hwang, S. J. (2024). Concept-skill Transferability-based Data Selection for Large Vision-Language Models. Conference on Empirical Methods in Natural Language Processing, 5060–5080.

Methods

- Multilayer activation features from a small reference VLM (e.g., TinyLLaVA-2B)
- concatenate MSA features across 5 layers

$$[\boldsymbol{z}_l^v, \boldsymbol{z}_l^t] = \mathrm{MSA}_l\left(\mathrm{LN}_l\left([\boldsymbol{x}_l^v, \boldsymbol{x}_l^t]\right)\right) + [\boldsymbol{x}_l^v, \boldsymbol{x}_l^t]$$

$$\boldsymbol{u}_l^v = \texttt{L2-Normalize}(\texttt{MeanPool}(\texttt{tanh}(\boldsymbol{z}_l^v))),$$
$$\boldsymbol{u}_l^t = \texttt{L2-Normalize}(\texttt{MeanPool}(\texttt{tanh}(\boldsymbol{z}_l^t))),$$

use tanh(.) to avoid activation extremes in LLMs

$$\boldsymbol{u}^m = [\boldsymbol{u}_{l_1}^v,\ \boldsymbol{u}_{l_1}^t,\ \ldots,\ \boldsymbol{u}_{l_M}^v,\ \boldsymbol{u}_{l_M}^t] \,/\, \sqrt{2M}$$

concatenate features from the small VLM's layers

- **Spherical k-means clustering** on u_m with large K (e.g., 10,000) to capture fine-grained concept-skill compositions

Lee, J., Li, B., & Hwang, S. J. (2024). Concept-skill Transferability-based Data Selection for Large Vision-Language Models. Conference on Empirical Methods in Natural Language Processing, 5060–5080.

Data selection criteria

Clusters closer in activation space transfer better to each other; *centroid cosine similarity can proxy transferability.*

Transferability proxy S_i: Average cosine similarity between cluster centroid e_i and other centroids.

$$S_i = \frac{1}{K_{\text{tgt}}} \sum_{j=1}^{K_{\text{tgt}}} \cos(\boldsymbol{e}_i, \boldsymbol{e}_j),$$

Density measure

- low D_i: diverse
- high D_i: dense

$$D_i = \frac{1}{|C_i|(|C_i|-1)} \sum_{p,q \in C_i, p \neq q} d(p,q)$$

Cluster-wise allocation probability

$$P_i \propto \exp(S_i/(\tau \tilde{D}_i))$$

Select sample from each cluster

$$N_{\text{core}} P_i$$

Intra-cluster selection via greedy MMD minimization

$$\mathrm{MMD}^2 = A(C_i, C_i) + A(C_i', C_i') - 2A(C_i, C_i'),$$

$$A(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d(p,q). \qquad (7)$$

Data selection algorithm

Algorithm 1 COINCIDE Data Selection Algorithm

Require: K: the number of clusters, N_{core}: target coreset size

Extract multimodal neuron activations $\boldsymbol{u}^m$ from the full dataset. ▷ Eq. 3

Cluster $\boldsymbol{u}^m$ into K clusters to form a set of clusters $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$.

Compute cluster transferability $S_i = \mathbb{E}_j\left(\cos(\boldsymbol{e}_i, \boldsymbol{e}_j)\right),\ i \in \{1, 2, \ldots, K\}$ ▷ Eq. 5

Compute cluster density $D_i = \mathbb{E}_{p,q \sim C_i}\left(d(p, q)\right),\ i \in \{1, 2, \ldots, K\}$ ▷ Eq. 6

Calculate cluster categorical distribution $P_i \propto \exp(S_i / (\tau D_i))$.

for $i = 1, 2, \ldots, K$ **do**

    i-th cluster empty coreset $\mathcal{C}'_i$.

    i-th cluster target sample size $N_{\text{core},i} = N_{\text{core}} P_i$.

    while $|\mathcal{C}'_i| < N_{\text{core},i}$ **do**

        $k = \underset{j \in \mathcal{C}_i \setminus \mathcal{C}'_i}{\text{argmin}}\ \text{MMD}^2\left(\mathcal{C}_i, \mathcal{C}'_i \cup \{j\}\right)$ ▷ Eq. 7

        $\mathcal{C}'_i \leftarrow \mathcal{C}'_i \cup \{k\}$

    end while

end for

return $\mathcal{C}'_1 \cup \mathcal{C}'_2 \cup \ldots \cup \mathcal{C}'_K$

Experiment Setup

- **Reference model:** TinyLLaVA-2B (default); also CLIP, TinyLLaVA-0.9B, LLaVA-1.5 7B considered.
- **Benchmarks**: VQAv2, GQA, VizWiz, SQA-I, TextVQA, POPE, MME, MMBench-en/cn, LLaVA-Bench, MM-Vet.
- **Metric:** average relative performance (Rel.) vs full finetuning.
- **Targets: LLaVA-1.5 7B** (default), 13B; LoRA, 1 epoch, 4×V100

- **Baselines:**
  - Random,
  - CLIP-Score
  - EL2N
  - Perplexity
  - SemDeDup
  - D2-Pruning
  - Self-Sup
  - Self-Filter.

COINCIDE enables efficient VIT with only 16.7–20% data, matching or exceeding full-dataset generalization while reducing wall-clock time by up to ~70%.

Result

- LLaVA-1.5 @20% data: COINCIDE Rel. 97.4% (close to full, best avg.; +1.6 pp over best baseline).
- Strong on 7/10 benchmarks.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Bench	Rel. (%)
Full-Finetune	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	100
Random	75.7	58.9	44.3	68.5	55.3	84.7	1483.0	62.2	54.8	65.0	95.8
CLIP-Score	73.4	51.4	43.0	65.0	54.7	85.3	1331.6	55.2	52.0	66.2	91.2
EL2N	76.2	58.7	43.7	65.5	53.0	84.3	1439.5	53.2	47.4	64.9	92.0
Perplexity	75.8	57.0	47.8	65.1	52.8	82.6	1341.4	52.0	45.8	68.3	91.6
SemDeDup	74.2	54.5	46.9	65.8	55.5	84.7	1376.9	52.2	48.5	**70.0**	92.6
D2-Pruning	73.0	58.4	41.9	**69.3**	51.8	85.7	1391.2	**65.7**	**57.6**	63.9	94.8
Self-Sup	74.9	59.5	46.0	67.8	49.3	83.5	1335.9	61.4	53.8	63.3	93.4
Self-Filter	73.7	58.3	**53.2**	61.4	52.9	83.8	1306.2	48.8	45.3	64.9	90.9
COINCIDE (Ours)	**76.5**	**59.8**	46.8	69.2	**55.6**	**86.1**	**1495.6**	63.1	54.5	67.3	**97.4**

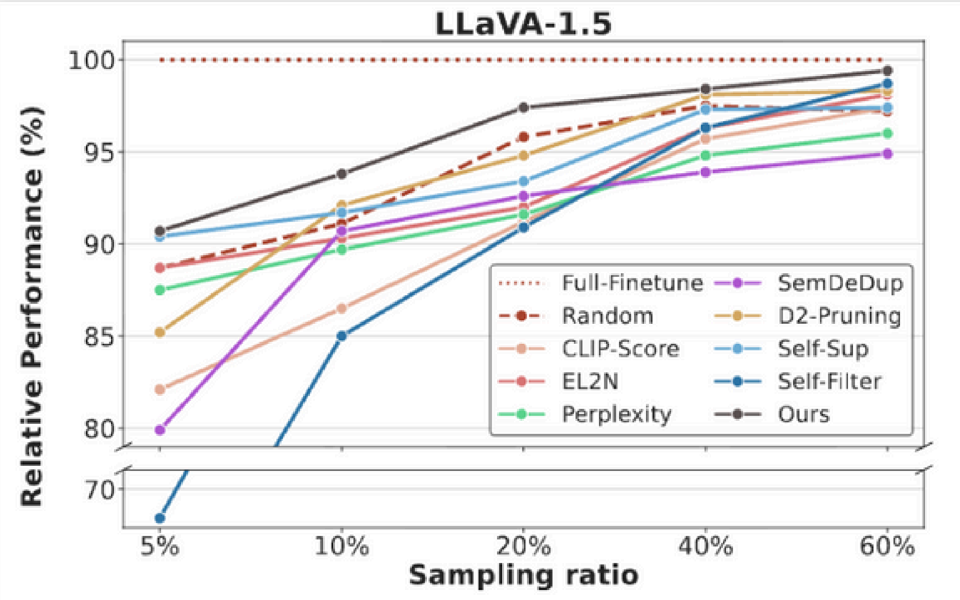


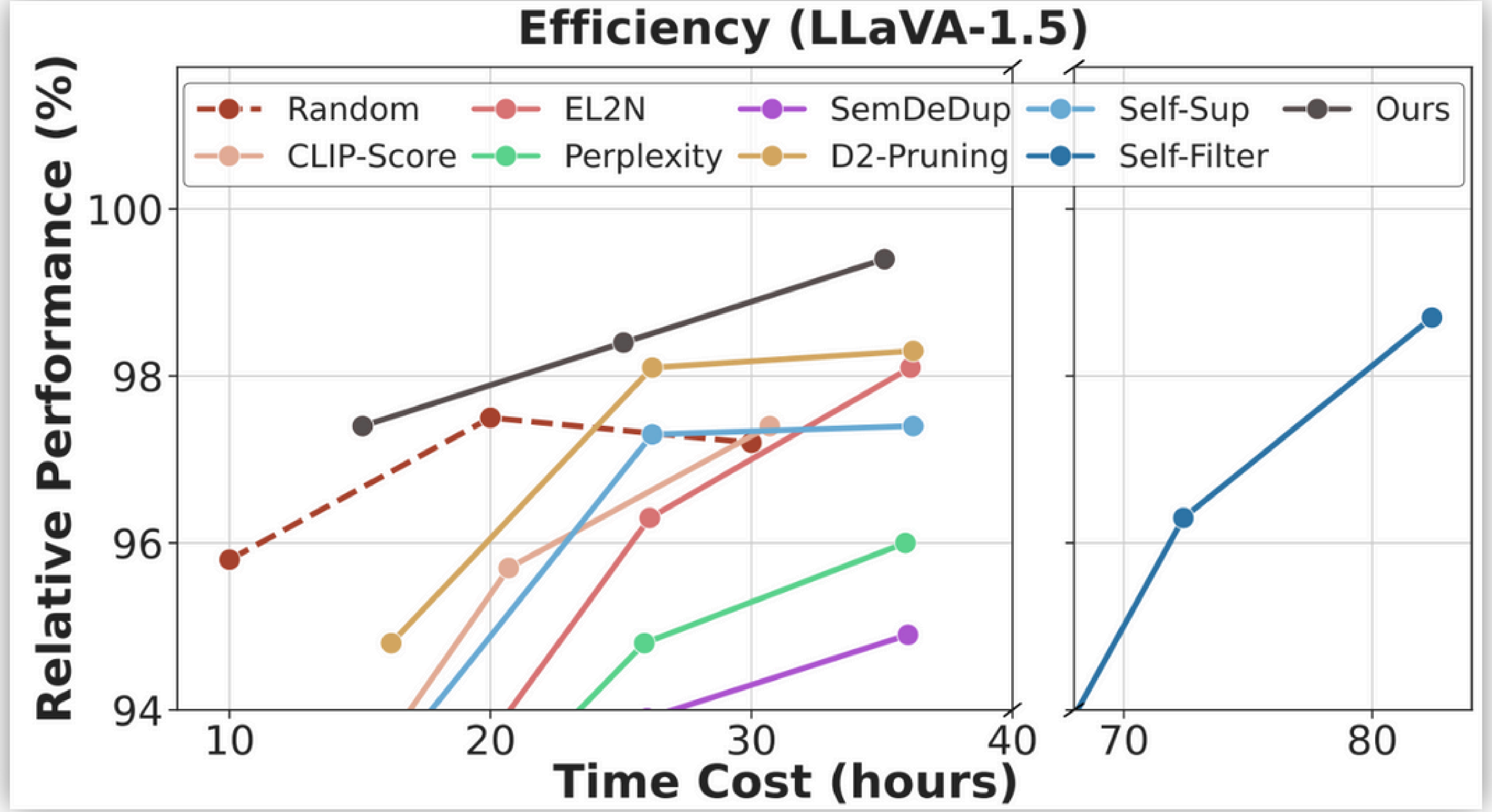
Figure 5: Average relative performances of all coreset selection techniques at different sampling ratios for the LLaVA-1.5 dataset.

Computational analysis

- No backward pass, only a small reference model
- k-means O(NK); cosine matrix O(K^2).

COINCIDE achieves 97.4%, 98.4%, and 99.4% relative performance with the wall-clock times of 15.1, 25.1, and 35.1 hours.

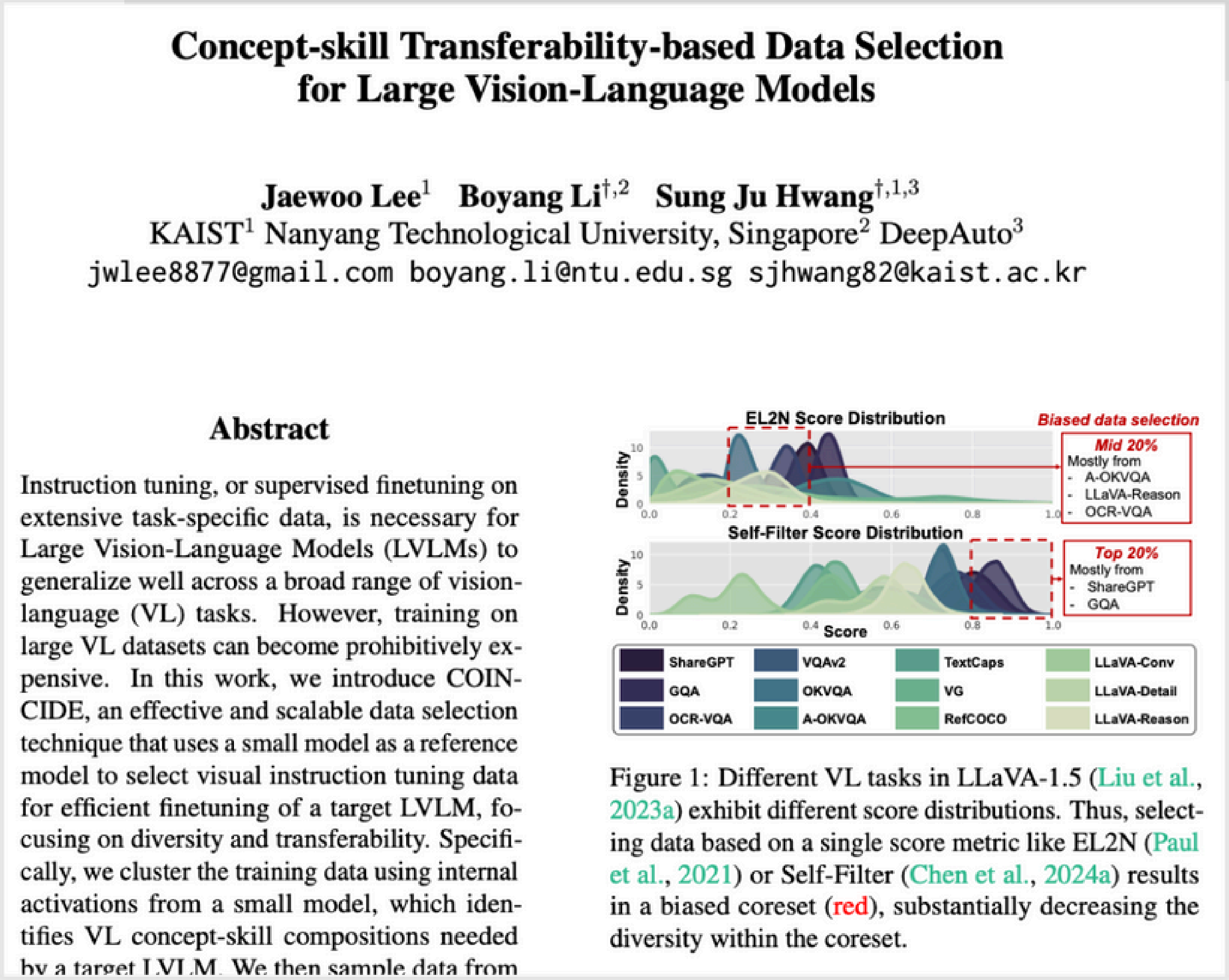
"Finetuning on all data takes 50 hours."



The wall-clock time cost of the entire pipeline of data selection and model finetuning ver□sus the average relative performance (Rel.) on the LLaVA-1.5 dataset.

Takeaways

- Diversity over concept-skill compositions is crucial for generalization; cluster-wise sampling ensures coverage beyond task-level sampling.
- Centroid cosine as a proxy unlocks cheap, effective transfer-aware allocation across clusters.
- Balancing transferability (S) and redundancy (D) yields efficient learning with fewer samples.
- Small reference models can reliably guide data selection for larger LVLMs.



Concept-skill Transferability-based Data Selection for Large Vision-Language Models

Jaewoo Lee[1] Boyang Li[†,2] Sung Ju Hwang[†,1,3]
KAIST[1] Nanyang Technological University, Singapore[2] DeepAuto[3]
jwlee8877@gmail.com boyang.li@ntu.edu.sg sjhwang82@kaist.ac.kr

Abstract

Instruction tuning, or supervised finetuning on extensive task-specific data, is necessary for Large Vision-Language Models (LVLMs) to generalize well across a broad range of vision-language (VL) tasks. However, training on large VL datasets can become prohibitively expensive. In this work, we introduce COINCIDE, an effective and scalable data selection technique that uses a small model as a reference model to select visual instruction tuning data for efficient finetuning of a target LVLM, focusing on diversity and transferability. Specifically, we cluster the training data using internal activations from a small model, which identifies VL concept-skill compositions needed by a target LVLM. We then sample data from

Figure 1: Different VL tasks in LLaVA-1.5 (Liu et al., 2023a) exhibit different score distributions. Thus, selecting data based on a single score metric like EL2N (Paul et al., 2021) or Self-Filter (Chen et al., 2024a) results in a biased coreset (red), substantially decreasing the diversity within the coreset.