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 these diverse clusters by considering their density and transferability, or the ability to transfer well to other concept-skill compositions. This approach ensures the diversity of these compositions, which is vital for LVLM generalization. Extensive experiments demonstrate that COINCIDE achieves superior performance and data selection efficiency against 8 strong baselines on two distinct datasets: LLaVA-1.5 and Vision-Flan. Using only 20% of the LLaVA-1.5 dataset, COINCIDE achieves performance comparable to the LVLM finetuned on the whole dataset, with 70% reduction of the wall-clock running time. On the Vision-Flan dataset, our method achieves superior results with only 16.7% of the training data. Our code is available at https://github.com/G-JWLee/COINCIDE_code.

1 Introduction

Large Vision-Language Models (LVLMs) (Zhu et al., 2023; Dai et al., 2023; Radford et al., 2021; Zhai et al., 2023) are often built by (1) pretraining on paired image-caption datasets and (2) subsequent finetuning on image-instruction data on diverse vision-language (VL) tasks. The second step, referred to as visual instruction tuning (VIT), substantially enhances multimodal instruction-following capabilities. To achieve broad generalization, recent works (Cha et al., 2023; Dong et al., 2024; Chen et al., 2024b; Li et al., 2024) integrate an increasing number of VL tasks into VIT.

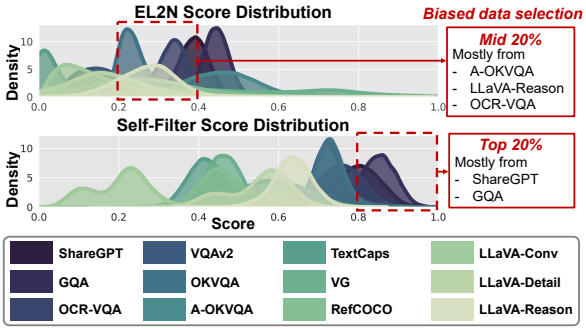


Figure 1: Different VL tasks in LLaVA-1.5 (Liu et al., 2023a) exhibit different score distributions. Thus, selecting data based on a single score metric like EL2N (Paul et al., 2021) or Self-Filter (Chen et al., 2024a) results in a biased coreset (red), substantially decreasing the diversity within the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

However, training on extensive VIT data incurs significant computational cost, making the process infeasible for small academic labs and individual researchers. Additionally, it is not clear if all the VIT data are necessary for good generalization, as different VL tasks have different abilities to transfer to downstream tasks (Tiong et al., 2024; Xi et al., 2023; Ostapenko et al., 2024).

In this paper, we investigate the selection of a coreset, a subset that approximates the performance of the full dataset, from large VIT datasets. Conventional coreset selection approaches (Marion et al., 2023; Zhou et al., 2023; Chen et al., 2023a) usually utilize a score metric to select training data. As VIT datasets are highly diverse and feature multiple data modes (Figure 1), data selection using any single metric would produce a coreset dominated by a few tasks. Figure 1 indicates that, selecting 20% of data from any part of the metric distribution of

† Equal advising

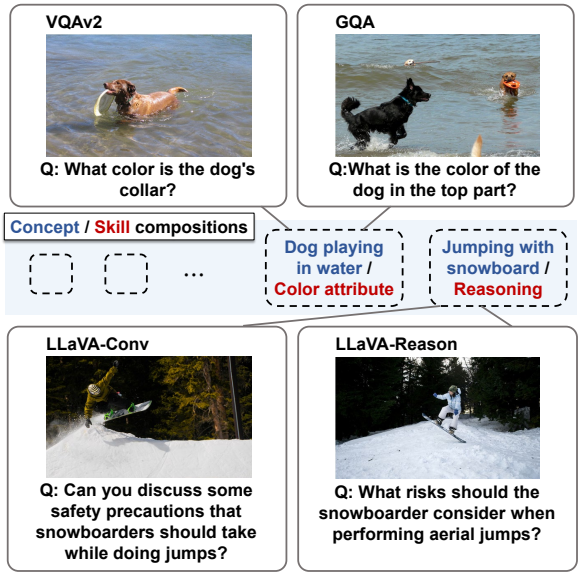


Figure 2: Different VL tasks (e.g., VQAv2 and GQA, LLaVA-Conv and LLaVA-Reason) share VL concept-skill compositions.

EL2N (Paul et al., 2021) or Self-Filter (Chen et al., 2024a) would exclude many data modes, which severely reduces the diversity of the selected coreset and harms generalization. As our experiments show (Table 1), this type of coreset selection degrades LVLM performance.

Our solution to the multitude of data modes is straightforward: we explicitly identify the modes by clustering the VIT data points using features from multiple layers in a small LVLM. Interestingly, we observe that the clusters thus identified roughly coincide with compositions of VL concepts and skills. For example, a concept could be street signs or trains on a railroad, while a skill could be OCR, recognizing color, or reasoning. Upon close inspection, we find that different VL tasks contain overlap over these concept-skill compositions. As exemplified in Figure 2, LLaVA-Conv and LLaVA-Reason contain questions about the risks of snowboard jumps, despite their separate focuses on multi-turn conversations and reasoning. This suggests sampling over the clusters would be more effective in enhancing the diversity of VL concept-skill compositions than sampling over datasets or tasks.

To this end, we introduce **CO**re **IN**struction **C**oncept-sk**I**ll **D**ata **E**lection (COINCIDE), which identifies VL concept-skill compositions through data clustering using activations from an off-the-shelf, small LVLM (Figure 3 Left). From each cluster, COINCIDE selects training data for a target LVLM by considering transferability (i.e., how well knowledge from each cluster can facilitate LVLM's learning in other clusters) and internal density of clusters (Figure 3 Right). Empirically, we find that transferability correlates well with cosine similarity among clusters. Based on the findings, we select more data points from more transferable clusters. Further, we sample fewer data points from denser clusters, as data points in dense clusters are likely redundant.

Another major challenge of coreset selection is its high computational cost. Existing techniques often require expensive steps like additional training (Du et al., 2023; Mekala et al., 2024; Chen et al., 2024a), gradient calculation (Xia et al., 2024; Liu et al., 2024), or the use of bigger and more advanced models (Chen et al., 2023a; Liu et al., 2023c). The time complexity and the assumption of larger models contradict the primary goal of coreset selection, which is to reduce the development cost of new models larger than existing ones. In comparison, COINCIDE assumes only a VLM (2B) smaller than the target LVLM (7B, 13B) and does not require any backward pass.

We validate the effectiveness of COINCIDE across a wide range of coreset selection scenarios using two distinct VIT datasets, LLaVA-1.5 (Liu et al., 2023a) and Vision-Flan (Xu et al., 2024). The experimental results demonstrate that our method achieves performance competitive with that of the LVLM finetuned with the full dataset, with 30% of time cost including the data selection and training. Our approach also achieves superior performance and efficiency compared to 8 strong baselines.

In summary, our contributions are as follows:

- We introduce COINCIDE, an efficient coreset selection pipeline for a target LVLM using an existing small reference model to cluster training data. Training on 16.7-20% data selected by COINCIDE achieves comparable performance to whole-dataset finetuning, leading to 70% wall-clock time reduction.
- We propose an efficient transferability calculation among clusters based on our novel observation of a positive correlation between cluster centroid similarity and cluster transferability.
- To enhance training efficacy, we prioritize samples from clusters with high transferability and low density, while still selecting a few samples from other clusters for diversity.

2 Related Work

Coreset Selection Coreset selection attempts to extract a subset of training data that functions comparably to the full training set. This technique is adopted for problems like active learning (Wei et al., 2015; Sener and Savarese, 2018), continual learning (Rebuffi et al., 2017; Aljundi et al., 2019), and data pruning (Pleiss et al., 2020; Paul et al., 2021). Recent works (Zhou et al., 2023; Xia et al., 2024) investigate coreset selection for instruction tuning of LLMs. Alpagasus (Chen et al., 2023a) uses ChatGPT (OpenAI, 2022) to rate the quality of instruction samples. S2L (Yang et al., 2024) leverages the training loss trajectory of smaller models to find optimal samples for training larger LLMs. DiverseEvol (Wu et al., 2023) utilizes the target model itself to iteratively choose beneficial data for the current training episode.

Coreset Selection for Visual Instruction Tuning Several very recent papers address the coreset selection problem for visual instruction tuning (Wei et al., 2023; Chen et al., 2024a; Liu et al., 2024). Self-Filter (Chen et al., 2024a) scores VIT data using a score-net trained along with the target LVLM. The concurrent work TIVE (Liu et al., 2024) employs gradient information from the target LVLM to compute task- and sample-level importance. Although effective, it demands considerable memory to store the high-dimensional gradient vectors. Moreover, these methods require backward passes, which are expensive due to the large training set. Both also overlook the diversity of selected data, which is vital for generalization. In contrast, our approach reduces wall-clock running time and considers both transferability and diversity.

VL Concept and Skill Discovery Concept discovery in neural networks is a key topic in interpretability research (Kim et al. 2018; FEL et al. 2023; Manning et al. 2020). Notably, Kowal et al. (2024) performs hierarchical clustering in layer-wise activation space. Tiong et al. (2024) attempts to identify latent skills underlying VL datasets. Michaud et al. (2023) performs spectral clustering to discover LLMs skills. Though these works provide inspiration, they are orthogonal to our work, whose main objective is to sample from data clusters rather than understanding existing neural networks. The only application of concept discovery we are aware of is by Gupta et al. (2017), showing consistent VL concepts improve transfer learning.

3 Method

We start by introducing the framework that utilizes neuron activations from a small LVLM to group VIT data into clusters, where each cluster comprises samples exhibiting a similar concept-skill composition (Section 3.2). Next, we conduct experiments to examine the correlation between the similarity of a cluster centroid to other centroids and the transferability of that cluster to others (Section 3.3). Based on our findings, we describe our data selection strategy, which performs cluster-wise sample selection by selecting different numbers of samples from clusters depending on their transferability and diversity (Section 3.4). The overall framework of our approach is illustrated in Figure 3.

3.1 Preliminaries

A modern LVLM typically consists of a visual encoder and an LLM, which are connected by intermediate network layers. The visual information is fed to the LLM as input (Dai et al. 2023; Liu et al. 2023b), or guides cross-attention (Alayrac et al. 2022). Here we focus on a transformer-based LLM that receives visual information as input tokens.

The l-th transformer layer receives the visual tokens $\boldsymbol{x}_l^v \in \mathbb{R}^{N_v \times D}$ and text tokens $\boldsymbol{x}_l^t \in \mathbb{R}^{N_t \times D}$, where N_v and N_t are the numbers of tokens, and D is the hidden dimension size. A transformer layer contains a multi-head self-attention (MSA) and a feed-forward network (FFN). For the purpose of this paper, we describe only MSA formally:

$$[\boldsymbol{z}_l^v, \boldsymbol{z}_l^t] = \mathsf{MSA}_l\left(\mathsf{LN}_l\left([\boldsymbol{x}_l^v, \boldsymbol{x}_l^t]\right)\right) + [\boldsymbol{x}_l^v, \boldsymbol{x}_l^t], \quad (1)$$

where $[\cdot, \cdot]$ denotes concatenation, LN_l denotes layer normalization, and $\boldsymbol{z}_l^v$ and $\boldsymbol{z}_l^t$ are output visual and text features from the l-th layer MSA, respectively.

3.2 Discovering Concept-Skill Compositions

An LVLM aims to learn about a large variety of visual-linguistic concepts and skills. Hence, it is important to automatically sort training data into concepts and skills, so that the coreset can provide sufficient coverage of these. Recent studies (Schwettmann et al., 2023; Pan et al., 2023; Gandelsman et al., 2024) reveal that the internal activations at various layers of LVLMs may encode different visual concepts.

To figure out which layer of the LVLM provides the best feature representation for visual concept and skill discovery, we perform a preliminary visualization study of TinyLLaVA-2B (Zhou et al.,

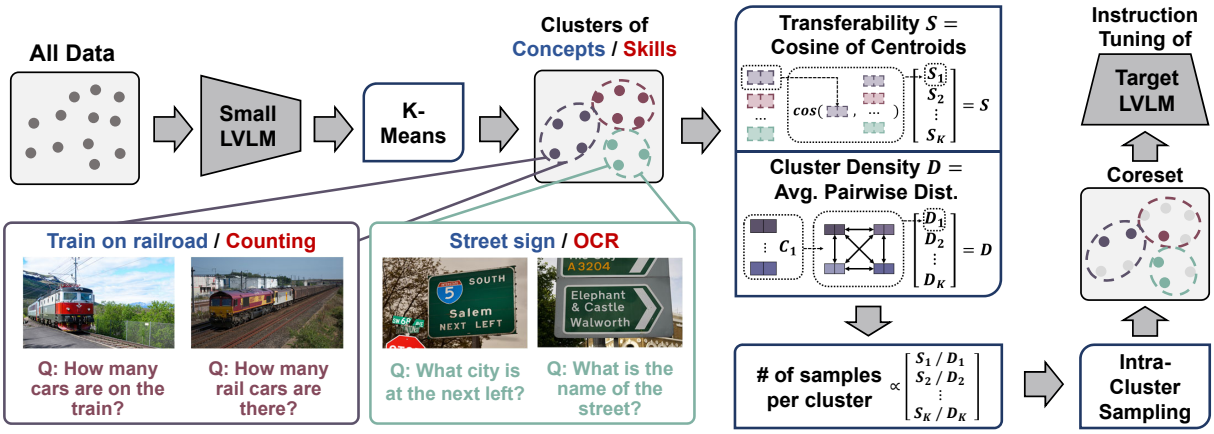


Figure 3: Illustration of COINCIDE. Our method utilizes a small LVLM to cluster visual instruction tuning data based on concept-skill compositions. We then assess the cluster transferability as the mean cosine similarity to other cluster centroids. We further compute the cluster density as the mean Gaussian kernel distance among all data pairs within the cluster. Using cluster transferability and density, COINCIDE determines the number of data to sample from each cluster and performs intra-cluster sampling. Finally, it combines all the selected samples from all the clusters to compose the final coreset.

2024). Given an image and a textual question, we visualize the image patches that contribute the most to the generation of the ground-truth answer. Using features from different layers highlights different image patches. Ideally, we can compare the visualization with human intuition and select the layer that agrees with human intuition the most. We provide detailed experimental procedures with some visualization results in Appendix B.

Perhaps surprisingly, we find that the best layer varies substantially according to the input. That is, the VL concepts and skills are distributed across different layers. Hence, for the clustering, we choose five layers spanning from the initial to top layers of the model to cover a wide range of concepts and skills and use the concatenation of their output as the feature vector of each data point.

We cluster VIT training data points using their feature vector from multiple layers of a small LVLM, called a reference model. We extract the features right after the MSA of the l-th layer (Eq. 1) and process them into unit-length vectors:

$$\begin{aligned} \boldsymbol{u}_l^v &= \texttt{L2-Normalize}(\texttt{MeanPool}(\texttt{tanh}(\boldsymbol{z}_l^v))), \\ \boldsymbol{u}_l^t &= \texttt{L2-Normalize}(\texttt{MeanPool}(\texttt{tanh}(\boldsymbol{z}_l^t))), \end{aligned} \quad (2)$$

where the mean-pooling is performed across the number of visual and text tokens, respectively. The hyperbolic tangent function, tanh, is necessary to reduce the impact of a few extreme activations, which are described by Sun et al. (2024). Without this step, these large values would dominate the feature vector and skew the clustering. After that, we concatenate features from the small LVLM's layers:

$$\boldsymbol{u}^m = [\boldsymbol{u}_{l_1}^v,\ \boldsymbol{u}_{l_1}^t,\ \ldots,\ \boldsymbol{u}_{l_M}^v,\ \boldsymbol{u}_{l_M}^t] \ / \ \sqrt{2M}, \quad (3)$$

where M denotes the number of layers where we extract the features, and the subscripts $l_1, \ldots l_M$ are the layer indices. The resultant $\boldsymbol{u}^m \in \mathbb{R}^{2M*D}$ is the final multimodal feature of the data point.

Then, we perform spherical k-means clustering on $\boldsymbol{u}^m$, yielding K clusters. To ensure the purity of clusters, we set K to a large number, such as $10,000$. Despite its simplicity, the k-means procedure runs in $O(NK)$ time for N data points, which is advantageous when both N and K are large. Other clustering techniques such as spectral clustering or affinity propagation are much more expensive. Qualitative analysis indicates the clusters coincide with concept-skill compositions. We provide visualization of the clusters in Appendix C.

3.3 Measuring Cluster Transferability

Empirical evidence shows that datasets differ in their ability to generalize to other datasets (Zamir et al., 2018; Achille et al., 2020). We hypothesize that (1) data clusters also have varying levels of transferability and (2) clusters close together in feature space transfer well to each other. If (1) is true, it would be beneficial to select data from highly transferable clusters. If (2) is true, we can use distance among clusters as a proxy for transferability.

We design an experiment to verify the hypotheses. Following Chen et al. (2023b), to measure

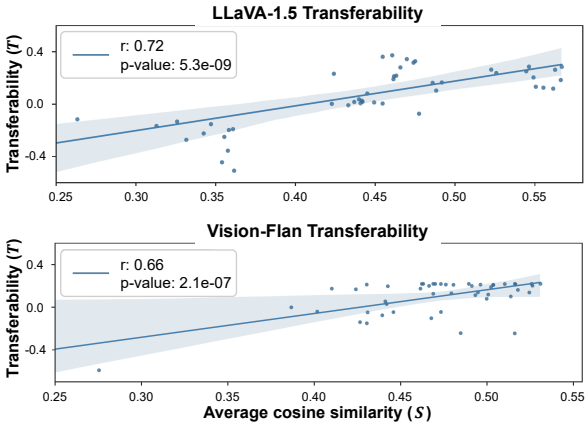


Figure 4: Correlation between cluster centroid similarity and transferability. We examine the correlations in the LLaVA 1.5 (Liu et al., 2023a) and Vision-Flan (Xu et al., 2024) datasets, with each point representing a source cluster. We report the Pearson correlation coefficient (r) and p-value.

transferability from cluster C_i to cluster C_j, we run two training sessions. In the first, we finetune an LVLM on the same number of samples, N_c, drawn from C_i and C_j respectively. In the second, we finetune on N_c samples from C_j only. After finetuning, both models are tested on unseen samples from C_j, yielding test losses $L_{i,j\to j}$ and $L_{j\to j}$. The difference $L_{j\to j} - L_{i,j\to j}$ can be seen as the degree by which C_i facilitates the learning of C_j. We aggregate over target clusters to compute the transferability of the source cluster C_i:

$$T_i = \frac{1}{K_{\text{tgt}}} \sum_{j=1}^{K_{\text{tgt}}} (L_{j\to j} - L_{i,j\to j}), \tag{4}$$

where K_{tgt} is the number of target clusters. Then, we compute the cosine similarity of the source cluster with the target clusters and average:

$$S_i = \frac{1}{K_{\text{tgt}}} \sum_{j=1}^{K_{\text{tgt}}} \cos(\boldsymbol{e}_i, \boldsymbol{e}_j), \tag{5}$$

where $\boldsymbol{e}_i$ is the cluster centroid of cluster C_i.

We compute the correlation between transferability T_i and average cosine similarity S_i over all possible pairings between 50 random source clusters and 50 random target clusters, and plot the results in Figure 4. We find that (1) clusters differ significantly in transfer power, and (2) S_i and T_i have a strong positive correlation (0.66-0.72), indicating that the cosine similarity among clusters can serve as an effective and inexpensive proxy for transferability. For K clusters, the time complexity of all cosine similarities is $O(K^2)$. Further studies of transferability are available in Appendix D.

3.4 Data Selection Criteria

In addition to transferability T_i and its proxy S_i, we consider the density of a cluster during the sampling process, as selecting too many data points from a dense cluster that contains many similar samples would create redundancy. Hence, we introduce a density measure D_i:

$$D_i = \frac{1}{|C_i|(|C_i|-1)} \sum_{p,q\in C_i, p\neq q} d(p,q), \tag{6}$$

where p and q are two distinct data points from cluster C_i, and $d(p,q) = \exp(-\|\boldsymbol{u}_p^m - \boldsymbol{u}_q^m\|^2)$ is the Gaussian kernel function with $\boldsymbol{u}_p^m$ and $\boldsymbol{u}_q^m$ being the multimodal neuron activations (Eq. 3) of data points p and q, respectively. The small D_i value indicates that the cluster C_i is highly diverse.

In order to create a coreset of N_{core} samples, we select from cluster C_i exactly $N_{\text{core}} P_i$ samples. Here, $P_i \propto \exp(S_i/(\tau D_i))$ is a categorical distribution and τ is a temperature hyperparameter. This approach enables us to select more samples from more transferable and less dense clusters to enhance training efficacy, while still selecting a few samples from other clusters to ensure diverse concept-skill compositions in the coreset.

From cluster C_i, we aim to select $N_{\text{core}} P_i$ samples that are representative of the original data distribution of C_i. We compute the distance between the original cluster C_i and the set of sampled data points C_i' as MMD^2, the squared maximum mean discrepancy, which is defined as:

$$\begin{aligned} \text{MMD}^2 &= A(C_i, C_i) + A(C_i', C_i') - 2A(C_i, C_i'), \\ A(C_i, C_j) &= \frac{1}{|C_i||C_j|} \sum_{p\in C_i, q\in C_j} d(p,q). \end{aligned} \tag{7}$$

We iteratively add samples from the cluster C_i to the sampled cluster C_i' that minimizes MMD^2 using greedy search (Kim et al., 2016). In the end, we combine all the selected samples from all the clusters to compose the final VIT coreset. The complete data selection algorithm is shown in Appendix G.

4 Experiments

4.1 Setup

Visual Instruction Tuning Datasets We conduct coreset selection on two distinct VIT datasets: LLaVA-1.5 (Liu et al., 2023a) and Vision-Flan (Xu

Table 1: Comparison of coreset selection techniques on the LLaVA-1.5 dataset. We finetune the models using coresets with a 20% sampling ratio and estimate performance on various multimodal evaluation benchmarks. The best and the second best results are in **bold** and underlined, respectively.

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**

et al., 2024). The LLaVA-1.5 dataset contains 665k VIT data from 12 different VL tasks. The Vision-Flan dataset comprises 191 VL tasks, each with approximately 1k expert-annotated VIT data points, totaling 186k samples.

Models for Training and Data Selection For the target LVLMs, we use the pre-trained LLaVA-1.5 model (Liu et al., 2023a) with a default size of 7B parameters unless otherwise specified. In all experiments, we train the models using LoRA (Hu et al., 2022) for one epoch, following the official finetuning hyperparameters specified in LLaVA-1.5. As a reference model, we use the TinyLLaVA-2B (Zhou et al., 2024), a small LVLM finetuned on the target VIT dataset, for efficient coreset selection for all methods unless otherwise specified. All experiments are conducted using 4 V100 GPUs.

Evaluation Benchmark To assess the generalization of finetuned LVLMs across diverse visual instructions, we evaluate the models on several widely adopted zero-shot multimodal evaluation benchmarks, including 1) visual question answering: VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018); 2) knowledge-grounded QA: ScienceQA (Lu et al., 2022); 3) Optical Character Recognition (OCR): TextVQA (Singh et al., 2019); 4) hallucination: POPE (Li et al., 2023); 5) multiple-choice: MME (Fu et al., 2023), MMBench (Liu et al., 2023d); 6) free-form generation: LLaVA-Bench (Liu et al., 2023b), MM-Vet (Yu et al., 2023). In all experiments, we follow the protocols outlined in LLaVA-1.5 and Vision-Flan to select evaluation benchmarks. Further explanations of these benchmarks are provided in Appendix A.

Since each evaluation benchmark has a different scale, we compute average relative performance, denoted as Rel., across benchmarks to assess the level of generalization. Each relative performance is derived from the formula: (model performance / full-finetuned performance) $\times$ 100%.

Baselines We compare our method with several coreset selection techniques: CLIP-Score, EL2N (Paul et al., 2021), Perplexity (Marion et al., 2023), SemDeDup (Abbas et al., 2023), D2-Pruning (Maharana et al., 2023), Self-Sup (Sorscher et al., 2022). We also compare with a recent VIT coreset selection method, Self-Filter (Chen et al., 2024a). We additionally report the results of *Random*, the model finetuned with the coreset collected by random sampling, and *Full-Finetune*, the model finetuned with the full VIT dataset. The details of the baseline methods are provided in Appendix A.

4.2 Results and Discussion

COINCIDE surpasses baselines on LLaVA-1.5. Table 1 presents model performance when we limit the coreset to 20% of the size of the LLaVA-1.5 VIT dataset. COINCIDE is either the best or a close second on 7 out of 10 benchmarks, including VQAv2, GQA, SQA-I, TextVQA, POPE, MME, and MMBench-en. On average, COINCIDE outperforms the best baseline by 1.6 percent points (pp) in relative performance.

Interestingly, all baselines perform worse than the random sampling on average relative performance, suggesting that they may be susceptible to the selection bias, which is discussed in the in-

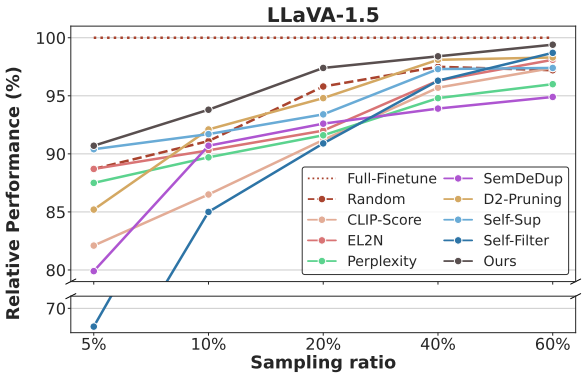


Figure 5: Average relative performances of all coreset selection techniques at different sampling ratios for the LLaVA-1.5 dataset.

Table 2: Comparison of coreset selection techniques on the Vision-Flan dataset. We finetune the models using coresets with a 16.7% sampling ratio and estimate performance on various multimodal evaluation benchmarks. The best and the second best results are in **bold** and underlined, respectively.

Method	MMBench-en	MME	MM-Vet	POPE	SQA-I	Rel. (%)
Full-Finetune	53.4	1287.5	25.6	84.2	61.3	100
Random	45.2	1122.3	26.1	82.5	60.9	94.2
CLIP-Score	34.3	687.6	26.6	72.6	61.8	81.7
EL2N	45.3	1082.9	23.9	82.1	60.6	91.7
Perplexity	39.3	1160.9	26.1	83.1	59.2	92.2
SemDeDup	42.1	1146.5	27.2	82.7	56.8	93.0
D2-Pruning	49.1	1052.4	27.0	82.5	**64.7**	96.5
Self-Sup	42.9	1012.2	23.5	80.8	60.0	88.9
Self-Filter	28.6	923.6	**30.0**	**83.3**	59.3	87.6
COINCIDE (Ours)	**56.7**	**1222.2**	26.2	81.9	63.8	**101.0**

troduction and illustrated in Figure 1. In contrast, COINCIDE considers the diversity of VL concept-skill compositions, demonstrating high generalization across a broad range of visual instructions. We further analyze the selection bias of the baselines and effectiveness of COINCIDE in Appendix E.

In Figure 5, we show the performance comparison across different coreset sizes as proportions of the original LLaVA-1.5 dataset. COINCIDE consistently outperforms other baselines across various sampling ratios, underscoring the effectiveness of our approach. COINCIDE also performs well on LLaVA-1.5-13B, as shown in Appendix F.1.

One Sixth of Vision-Flan selected by COINCIDE outperforms full dataset. We further evaluate the coreset selection techniques on the Vision-Flan VIT dataset (Xu et al., 2024) and show the results in Table 2. COINCIDE exceeds the performance of the model finetuned on the whole Vision-Flan data by 1.0 pp and the performance of the best baseline by 4.5 pp, using a selected subset 16.7% (1/6) of its size. Further, as illustrated in Figure 6, COINCIDE maintains consistently high performance across several sampling rates.

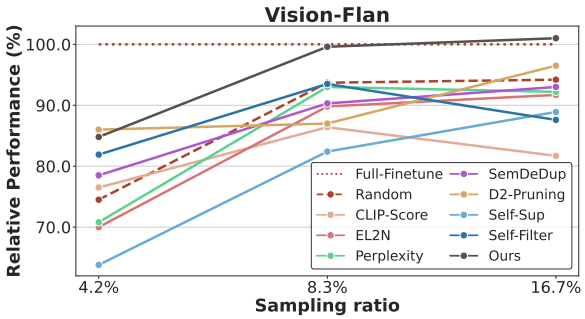


Figure 6: Average relative performances of all coreset selection techniques at different sampling ratios for the Vision-Flan dataset.

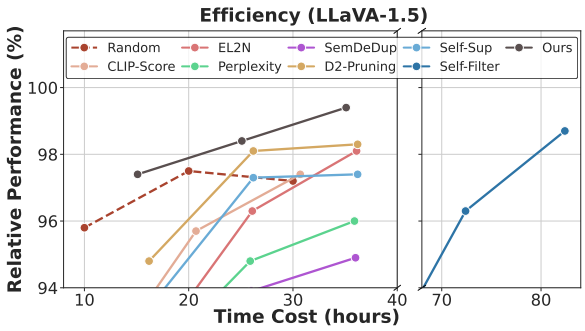


Figure 7: Comparison of coreset selection techniques on average relative performance and wall-clock time cost. The wall-clock time cost includes both the data selection and finetuning of the target LVLM. The time cost is measured in hours of running time on a computing node with 4× V100 GPUs.

We note that Vision-Flan, with its 191 VL tasks, is much more diverse than the LLaVA-1.5 dataset of 12 tasks. The stronger performance of COINCIDE on the Vision-Flan suggests that COINCIDE algorithm is well adapted to the use case of visual instruction tuning, which is increasingly performed on larger and more diverse sets of tasks.

Another curious phenomenon is that several baselines, including CLIP-Score, Perplexity, and Self-Filter, experience performance declines as the sampling ratio increases in Figure 6. A similar trend is observed in the random baseline in Figure 5. This underscores the importance of deliberate coreset selection, as merely increasing the dataset size does not guarantee improved LVLM capabilities.

Table 3: Ablation studies of COINCIDE. **(a)** Effect of different reference models. The time cost includes both the data selection and finetuning of the target LVLM and is measured in hours of running time on a computing node with 4× V100 GPUs. **(b)** Ablation on data selection criteria of our approach, transferability (S) and density (D). **(c)** The performance of different intra-cluster sampling strategies across various coreset sizes.

(a) Reference Model

Model (# params)	Time (hours)	Rel. (%)
CLIP (0.4B)	**10.9**	94.2
TinyLLaVA (0.9B)	12.2	96.3
TinyLLaVA (2B)	15.3	**97.4**
LLaVA-1.5 (7B)	20.7	97.1

(b) Key Components

Method	S	D	Rel. (%)
Random	–	–	95.8
COINCIDE (Ours)	–	–	94.4
	✓	–	95.9
	–	✓	94.7
	✓	✓	**97.4**

(c) Intra-Cluster Sampling methods

Intra-Cluster Sampling	Sampling ratio				
	5%	10%	20%	40%	60%
Random-select	90.1	**94.3**	**97.5**	97.7	98.3
Nearest-to-centroid	**91.9**	**94.3**	96.7	**99.1**	98.4
Greedy-MMD^2-minimize	90.7	93.8	**97.4**	98.4	**99.4**

COINCIDE provides wall-clock training time reduction and is Pareto superior. In Figure 7, we plot the wall-clock time cost of the entire pipeline of data selection and model finetuning versus the average relative performance (Rel.) on the LLaVA-1.5 dataset. 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, respectively. In contrast, finetuning on all data takes 50 hours.

We observe that COINCIDE provides Pareto superior solutions to all baselines. This is mainly due to the excellent time complexity of COINCIDE, which is linear to the number of training data points. Moreover, our method discovers the transferability among clusters at a low computational cost. It requires only cosine similarity calculations, with a time complexity quadratic to the number of clusters. Hence, COINCIDE provides a scalable data selection procedure.

COINCIDE also utilizes neuron activations from intermediate layers of the small reference model rather than the final outputs, avoiding complete forward passes like other baselines. Additionally, COINCIDE does not require training of additional networks that score data points, like Self-Filter. Neither does it require backward passes like the concurrent work TIVE (Liu et al., 2024). The combination of all these factors leads to an efficient solution to coreset selection.

4.3 Further Analysis and Ablation

Alternative Reference Models We analyze the effects of different reference models, which are the models used to extract features for clustering and cosine similarity. We compare four models, CLIP, TinyLLaVA-0.9B, TinyLLaVA-2B, and LLaVA-1.5-7B, and report the time cost of the entire coreset selection pipeline and average relative performance in Table 3 (a). We observe that CLIP performs the worst whereas TinyLLaVA-2B performs the best with reasonable time cost in data selection. However, the differences between TinyLLaVA-0.9B, TinyLLaVA-2B, and LLaVA-1.5-7B are small. We conclude that a well-trained small model can serve effectively as a reference model in coreset selection for a target LVLM. We also examine the robustness of COINCIDE when the reference model is finetuned on a different VIT dataset, which is detailed in Appendix F.2.

Ablation on Data Selection Criteria To validate our coreset selection method, we conduct ablation studies on the two data selection criteria, transferability and density, as summarized in Table 3 (b). In the first ablation, without using either criterion, we simply select the same number of samples from each cluster. This results in inferior performance, which suggests that naive stratified sampling from the clusters is not sufficient, possibly due to the heterogeneous nature of the clusters. In the second ablation, number of samples from each cluster is proportional to the transferability of the cluster, leading to a 1.5 percentage point (pp) increase. The third ablation selects number of samples inversely proportional to density, yielding a modest enhancement of 0.3 pp. Finally, combining both transferability and density provides a sizeable increase of 3.0 pp, demonstrating that the two selection criteria are complementary to each other.

Intra-cluster Selection Criteria COINCIDE selects samples within a cluster by minimizing MMD^2. We examine the effects of two alternative techniques, random selection and selecting samples closest to the centroids. As shown in Table 3 (c), in small coresets, samples closest to the centroids,

which are probably not outliers or hard samples, lead to high performance. In contrast, under high sampling ratios (i.e., large coresets), selecting diverse data using the MMD^2 metric leads to high performance. This is reminiscent of the finding of Sorscher et al. (2022) that easy samples are beneficial when the sampling ratio is small, whereas hard samples are advantageous when the sampling ratio is large. Overall, COINCIDE is robust to the choice of intra-cluster sampling, but adapting the intra-cluster sampling method to the sampling ratio can enhance the effectiveness of our approach.

5 Conclusion

In this paper, we introduce COINCIDE, a cluster-level data selection technique for efficient visual instruction tuning of Large Vision-Language Models. We demonstrate that clustering based on internal activations from a small model can represent visual-linguistic concept-skill compositions shared among diverse tasks in visual instruction tuning datasets. Additionally, our empirical investigation validates a strong positive correlation between cosine similarity and transferability among clusters. Based on the transferability and density of clusters, COINCIDE selects more samples from more transferable and less dense clusters to enhance training efficacy, while preserving the diversity of concept-skill compositions within the coreset to ensure better model generalization ability. Comprehensive experiments on the LLaVA-1.5 and Vision-Flan datasets demonstrate that our method outperforms baselines across several benchmarks with the lowest data selection cost, showcasing its effectiveness and efficiency. The success of COINCIDE suggests redundancy in popular VIT datasets and underscores the importance of a thorough understanding of data in training LVLMs.

Limitations

In our experiments, we observe that VL concept-skill compositions are shared across various VL tasks and identify VL concept-skill compositions that transfer well to others. However, after identifying these compositions and performing coreset selection, we finetune the target LVLMs by randomly selecting samples from the coreset. Recognizing the growing research attention on the importance of training order in LLM instruction tuning, we believe that considering the training order for LVLMs is crucial to enhance efficiency in visual instruction tuning. In future research, we aim to develop a curriculum learning algorithm that automatically determines the optimal training order based on the identified VL concept-skill compositions to further reduce the development cost of a new model.

Additionally, we assess whether the data with similar concept-skill compositions are concentrated well on the clusters through human inspection. Therefore, further investigation should be conducted to quantitatively evaluate the clustering of data with similar concept-skill compositions, which may enable accurate identification of VL concept-skill compositions and accurate quantification of their transferability.

Ethics Statement

In this work, we use publicly available visual instruction tuning datasets for coreset selection to enable easy replication. However, some data in the datasets contain erroneous answers about the visual content or images that do not clearly connect with the provided answers. Finetuning Large Vision-Language Models (LVLMs) with such data may lead to the generation of erroneous interpretations of images or hallucinations. This may pose an ethical issue for LVLM deployment in the real world. However, current coreset selection techniques, including ours, do not address hallucination in their selection processes. This motivates further research in coreset selection to identify visual instruction tuning data that minimizes hallucinations, aiming to build more reliable and trustworthy LVLMs.

Acknowledgements

Jaewoo Lee and Sung Ju Hwang are supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00256259) and a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (No. RS-2024-12345678). Boyang Li is supported by the Nanyang Associate Professorship and Fellowship (NRF-NRFF13-2021-0006) of the National Research Foundation, Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. 2020. The information complexity of learning tasks, their structure and their distance. *arXiv Preprint 1904.03292*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *arXiv Preprint 2204.14198*.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal LLM. *arXiv preprint arXiv:2312.06742*.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023a. Alpagasus: Training A better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue WANG, Ce Zhang, Frederic Sala, and Christopher Ré. 2023b. Skill-it! a data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024a. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye1, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *CoRR*, abs/2311.15653.

Thomas FEL, Victor Boutin, Louis Béthune, Remi Cadene, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. 2023. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2024. Interpreting clip's image representation via text-based decomposition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanmay Gupta, Kevin Shih, Saurabh Singh, and Derek Hoiem. 2017. Aligned image-word representations improve inductive transfer across vision-language tasks. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering

visual questions from blind people. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the International Conference on Machine Learning (ICML)*.

Matthew Kowal, Richard P Wildes, and Konstantinos G Derpanis. 2024. Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023c. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2023. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. *arXiv preprint arXiv:2402.10430*.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. 2023. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Oleksiy Ostapenko, Zhan Su, Edoardo Maria Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas Caccia, and Alessandro Sordoni. 2024. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*.

Haowen Pan, Yixin Cao, Xiaozhi Wang, and Xun Yang. 2023. Finding and editing multi-modal neurons in pre-trained transformer. *arXiv preprint arXiv:2311.07470*.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In

Advances in Neural Information Processing Systems (NeurIPS).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. 2023. Multimodal neurons in pretrained text-only transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gabriela Ben Melech Stan, Raanan Y. Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Anthony Meng Huat Tiong, Junqi Zhao, Boyang Li, Junnan Li, Steven C. H. Hoi, and Caiming Xiong. 2024. What are we measuring when we evaluate large vision-language models? an analysis of latent factors and biases. In *North American Chapter of the Association for Computational Linguistics*.

Kai Wei, Rishabh K. Iyer, and Jeff A. Bilmes. 2015. Submodularity in data subset selection and active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.

Zhiheng Xi, Rui Zheng, Yuansen Zhang, Xuanjing Huang, Zhongyu Wei, Minlong Peng, Mingming Sun, Qi Zhang, and Tao Gui. 2023. Connectivity patterns are task embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*.

Yu Yang, Siddhartha Mishra, Jeffrey N. Chiang, and Baharan Mirzasoleiman. 2024. Smalltolarge (S2L): scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *arXiv preprint arXiv:2403.07384*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Details of Experimental Setups

Evaluation Benchmark We provide in-depth explanations of the multimodal evaluation benchmarks used in our experiments. (1) VQAv2 (Goyal et al., 2017) evaluates the ability to understand and reason about general visual content by answering open-ended questions based on images. (2) GQA (Hudson and Manning, 2019) assesses compositional reasoning and understanding skills, requiring models to understand relationships and attributes of objects within images. (3) Vizwiz (Gurari et al., 2018) is designed to evaluate the model's ability to cope with real-world visual impairments. (4) ScienceQA-Image (SQA-I) (Lu et al., 2022) tests the model's science-related reasoning and visual understanding of images. (5) TextVQA (Singh et al., 2019) specifically targets text in images, assessing the Optical Character Recognition (OCR) ability of models. (6) POPE (Li et al., 2023) measures object hallucination in models. (7) MME (Fu et al., 2023) contains binary choice questions designed to evaluate perception and cognition abilities through 14 subtasks. (8) MMBench (Liu et al., 2023d) evaluates various abilities of models, covering object detection, text recognition, relation reasoning, etc., using tests conducted in English (en) or Chinese (cn). (9) LLaVA-Bench (Liu et al., 2023a) is specifically designed for evaluating models on visual instruction-following and chat ability. (10) MM-Vet (Yu et al., 2023) measures VL capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math.

Baselines In this section, we provide a more detailed explanation of the baselines. The hyperparameters for baselines in our experiments are summarized in Table 4.

- **CLIP-Score** utilizes the CLIP (Radford et al., 2021) model to assess the alignment between images and their instructions. For our study, we select VIT data with the highest CLIP scores.
- **EL2N** (Paul et al., 2021) estimates sample quality using the Error L2-Norm score, defined as $\mathbb{E}[||p(x) - y||_2]$. Here, $p(\cdot)$ represents the reference model, x is the input, and y is the ground-truth label. This metric calculates the average L2 distance between the model's predictions and the ground-truth labels for text tokens.
- **Perplexity** (Marion et al., 2023) measures the average negative log-likelihood of the next token prediction, defined as $\exp(-\mathbb{E}[\log p(x)])$. This metric assesses the uncertainty in the model's predictions. For both EL2N and Perplexity, we select data from the middle score distribution, as this range has been shown to perform best in prior research (Marion et al., 2023).
- **SemDeDup** (Abbas et al., 2023) removes semantically duplicated data by clustering the output embeddings of the last token from the reference model's final layer. This helps in reducing redundancy in the selected coreset.
- **D2-Pruning** (Maharana et al., 2023) represents the dataset as a graph where nodes represent sample difficulty and edges represent distances between samples. It actively uses the graph to preserve diversity in the coreset. We use the AUM (Pleiss et al., 2020) score to indicate difficulty, defined as $p_y(x) - \max_{i \neq y} p_i(x)$, where $p_y(x)$ is the prediction value for the ground-truth label, and $\max_{i \neq y} p_i(x)$ is the highest prediction value for any non-ground-truth label. For the distances between samples, we calculate the L2 distance between averaged output embeddings from the last layer tokens of the reference model.
- **Self-Sup** (Sorscher et al., 2022) clusters the data using the averaged output embeddings from the last layer tokens of the reference model. It scores data based on their distance to cluster centroids, selecting those the most likely to be prototypical.
- **Self-Filter** (Chen et al., 2024a) is a recent VIT coreset selection method that was originally applied to the LLaVA-158k VIT dataset (Liu et al., 2023b), which consists of only three VL tasks. It finetunes the score-net along with the target LVLM on the full dataset to serve as a reference model for scoring and filtering VIT data. We use the version that additionally incorporates both CLIP scores and CLIP features since it ensures enhanced performance and efficiency.

Table 4: Hyperparameter configurations.

Method	LLaVA-1.5	Vision-Flan
CLIP-Score	high score selected	high score selected
EL2N	medium score selected	medium score selected
Perplexity	medium score selected	medium score selected
SemDeDup	$\boldsymbol{K}$: 10,000	$\boldsymbol{K}$: 5,000
D2-Pruning	$\boldsymbol{k}$: 5, γ_r : 0.4, γ_f : 1.0	$\boldsymbol{k}$: 5, γ_r : 0.4, γ_f : 1.0
Self-Sup	$\boldsymbol{K}$: 10,000	$\boldsymbol{K}$: 5,000
Self-Filter	$\boldsymbol{k}$: 10, γ : 1	$\boldsymbol{k}$: 10, γ : 1
COINCIDE (Ours)	$\boldsymbol{K}$: 10,000, τ : 0.1	$\boldsymbol{K}$: 5,000, τ : 0.1

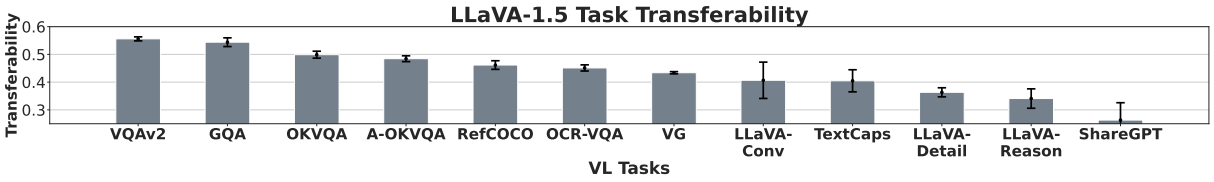


Figure 8: Task-wise transferability. We group the VIT data based on task names and then report the average cluster transferability of each group.

B Visualizing LVLM Skills with Relevancy Maps

In our method, we extract neuron activations from various layers (Eq. 2) to represent the concepts and skills of each VIT data. In this approach, we hypothesize that distinct layers represent distinct concepts and skills of the LVLM. To support this assumption, we compute relevancy maps (Chefer et al., 2021) following the approach outlined in Stan et al. (2024). The relevancy maps help us understand the model's final output by highlighting the most contributing parts of the input for each layer. Given the target output token $\boldsymbol{y_t}$ and the attention map $\boldsymbol{A}_l \in \mathbb{R}^{h\times(N_v+N_l)\times(N_v+N_l)}$ of the $\boldsymbol{l}$-th layer, where h is the head dimension of the attention, the relevancy map $\boldsymbol{R}$ is computed as follows:

$$\bar{\boldsymbol{A}}_l = \mathbb{E}_h[\nabla \boldsymbol{A}_l \odot \boldsymbol{A}_l],\ \nabla \boldsymbol{A}_l = \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{A}_l}, \\ \boldsymbol{R} = \boldsymbol{R} + \bar{\boldsymbol{A}}_l \cdot \boldsymbol{R}, \quad \text{for } l \in \{1, 2, \ldots, L\}, \tag{8}$$

where $\odot$ denotes the Hadamard product and L is the total number of layers in the LVLM. In order to investigate the contribution of each layer to the final output, we visualize the image regions related to the output token through the visual relevancy map computed from each layer. Specifically, we consider the row of $\bar{\boldsymbol{A}}_l \cdot \boldsymbol{R}$ corresponding to the output token. Then, we extract the visual token parts of the row to yield the visual relevancy map.

For the investigation, we inspect the 4th, 8th, 12th, 16th, and 20th layers of the TinyLLaVA-2B (Zhou et al., 2024) model and identify the layer that activates the most relevant visual regions. The findings in Figure 10 reveal that (1) the most relevant layer varies according to the concept-skill composition and (2) the most relevant layer is the same across diverse VIT data when the data shares a similar concept-skill composition. This supports our assumption that different layers contribute to distinct concepts and skills, allowing neuron activations from various layers to effectively group VIT data by their concept-skill composition.

C Concept-Skill Clustering Visualization

We visualize the clustering results of the gathered VIT data. The results are illustrated in Figure 11. We observe that most clusters contain VIT data that encode similar concept-skill compositions. For instance, the first group in Figure 11 consists of samples requiring OCR and counting abilities to solve visual queries involving images with store signs. The second group features images of people waiting for public transportation and multiple-choice questions that require visual recognition and reasoning abilities. The third group shows a cluster of samples with images of people in suits and queries focusing on object localization and generating captions for given bounding boxes. Lastly, the bottom group includes images exhibiting children with animals and requiring the ability to reason about the educational benefits that the children might gain from interacting with the animals.

D In-Depth Analysis on Concept-Skill Composition Transferability

D.1 Task-wise Transferability

To further understand transferability, we calculate the transferability of LLaVA-1.5 tasks by averaging the cluster transferability of VIT data. We show the results in Figure 8. We observe that VQA tasks, including VQAv2, GQA, OKVQA, and A-OKVQA, contain VIT data that transfers well to other data. In contrast, GPT-generated conversational tasks, including LLaVA-Conv, LLaVA-Detail, LLaVA-Rason, and ShareGPT, exhibit low transferability. This corresponds to the findings of Tiong et al. (2024) that VQA tasks are effective for finetuning LVLMs. This alignment supports the efficacy of our approach in discovering the fine-grained concept-skill compositions and their transferability. We hypothesize that the high transferability of the VQA tasks is because these tasks mostly require abilities close to the fine-grained VL concepts and skills that can be shared with other tasks, as described in Figure 2, unlike more complex tasks.

Table 5: Transferring to the larger target model. We validate if the coresets selected from TinyLLaVA-2B are transferable to LLaVA-1.5-13B finetining. We train the LLaVA-1.5-13B using coresets with 20% sampling ratio and estimate performance on various multimodal benchmarks. The best and the second best results are highlighted in **bold** and underline, respectively.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Wild	Rel. (%)
Full-Finetune	80.0	63.3	58.9	71.2	60.2	86.7	1541.7	68.5	61.5	69.5	100
Random	76.7	**60.5**	48.0	68.8	57.7	84.8	1484.9	62.8	55.2	68.6	94.0
CLIP-Score	75.3	52.6	42.2	69.7	57.3	85.4	1426.3	60.4	54.0	68.1	90.7
EL2N	77.2	59.6	**54.8**	69.9	56.1	84.1	**1531.0**	59.3	52.3	65.8	93.8
Perplexity	77.0	58.5	48.2	68.7	54.8	83.1	1508.8	57.5	50.3	68.7	91.6
SemDeDup	75.6	57.5	48.3	**70.5**	57.7	85.3	1397.6	59.0	51.1	68.7	91.9
D2-Pruning	73.9	**60.5**	49.8	70.4	55.2	84.9	1463.0	**67.3**	**59.9**	66.5	94.7
Self-Sup	76.3	**60.5**	50.0	70.2	52.7	85.4	1463.8	63.7	57.6	64.9	93.6
Self-Filter	75.0	59.8	48.6	69.5	55.8	84.5	1446.9	58.8	51.8	**69.1**	92.2
COINCIDE (Ours)	**77.8**	60.4	51.6	70.0	**58.6**	**87.1**	1516.8	64.0	57.7	67.4	**95.9**

D.2 Concept-Skill with High Transferability

In Figure 12, we visualize concept-skill compositions having the highest transferability for various VL task types. We define the VL task type of a cluster based on the task name associated with most of the cluster's data (e.g., VQAv2, GQA). Interestingly, GQA and LLaVA-Conv share a similar concept-skill composition as their most transferable concept-skill composition. This suggests that the transferability of VL concept-skill composition might be consistent across different VL tasks.

D.3 Concept-Skill as Latent Factor of LVLM

We conduct an ablation study to verify if data clusters from different VL task types have high transferability with each other when they share a similar concept-skill composition. In this study, we select two clusters from different VL task types with a similar concept-skill composition (second and fourth groups in Figure 12), using the first cluster as the source and the second cluster as the target. Additionally, we employ 49 randomly selected source clusters and measure transferability from the source clusters to the target cluster (Eq. 4). The source cluster, sharing a similar concept-skill composition with the target, ranks in the top 5 of the 50 source clusters in terms of test loss gain, exhibiting high transferability to the target cluster. This suggests that concept-skill compositions resemble fine-grained latent factors that constitute LVLM abilities. Thus, these fine-grained VL concepts and skills must be considered to effectively reduce data redundancy and build a well-generalized LVLM.

Table 7: Impact of a reference model training dataset. We use TinyLLaVA-2B finetuned on the LLaVA-1.5 dataset as a reference model to collect coresets from the Vision-Flan dataset with 16.7% sampling ratio. The best and the second best results are highlighted in **bold** and underline, respectively.

Method	MMBench-en	MME	MM-Vet	POPE	SQA-I	Rel. (%)
Full-Finetune	53.4	1287.5	25.6	84.2	61.3	100
EL2N	41.8	1082.0	23.9	82.6	61.7	90.9
Perplexity	45.7	1001.7	26.1	81.9	**64.8**	93.7
SemDeDup	46.8	1129.7	**27.2**	82.5	64.3	96.9
D2-Pruning	48.1	**1143.0**	27.0	83.4	63.1	97.3
Self-Sup	47.1	1084.6	23.5	81.7	63.5	93
COINCIDE (Ours)	**51.7**	1139.0	26.9	**84.0**	64.5	**99.1**

E Concept-Skill Diversity within Coresets

Our method selects data from various clusters to ensure a high diversity of VL concept-skill compositions within the coreset. To demonstrate the efficacy of our method, we compare the diversity within the coreset by our method with those by the baseline methods. Specifically, we use the 191 tasks from the Vision-Flan dataset as proxies for different concept-skill compositions, as there are no ground-truth compositions. We then count the number of selected samples for each task. The results, summarized in Figure 13, indicate that baseline methods select most data from only a few tasks, leading to biased selection and undermining LVLM generalization. This bias explains why most baselines perform worse than random sampling in our experiments. In contrast, our method achieves a more balanced selection across the various tasks.

Figure 9: Hyperparameter search. We examine the effect of the temperature (τ) and the number of clusters (K).

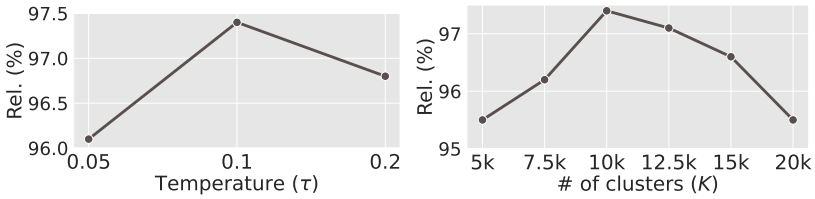


Table 6: We investigate the impact of various representations of multimodal neuron activation.

Neuron Activation	Rel. (%)
Boolean	95.7
Last layer	96.5
MSA layers	**97.4**
FFN layers	96.0

F Additional Experimental Results

F.1 Transfering to Larger Target Model

We evaluate the performance of the larger target model (LLaVA-1.5-13B) finetuned on coresets gathered by the small LVLM (TinyLLaVA-2B). Table 5 summarizes the performances across various benchmarks. The results demonstrate the effectiveness of our method in selecting a coreset that can be successfully transferred to the larger target model.

F.2 Robustness of Reference Model

We investigate the robustness of our method when the reference model is finetuned on a VIT dataset different from a target VIT dataset. To this end, we use the TinyLLaVA-2B finetuned on the LLaVA-1.5 VIT dataset, to perform coreset selection from the Vision-Flan dataset. The results are summarized in Table 7. COINCIDE continues to show performance comparable to full-finetuning while outperforming other baseline methods.

F.3 Hyperparameters

We conduct ablation studies on hyperparameters of our method, which include the number of clusters ($\boldsymbol{K}$) and the temperature ($\boldsymbol{\tau}$). The results, summarized in Figure 9, reveal that a sufficiently large number of clusters is essential to ensure cluster purity and diversity of VL concept-skill compositions, ensuring effective representation of the compositions and enhancing the generalization ability of LVLM. Furthermore, we find that setting the temperature too low leads to a biased coreset selection, as most samples are then selected from a few clusters. This undermines the diversity within the coreset, leading to a decline in overall performance.

F.4 Multimodal Neuron Activation

We further analyze the impact of different multimodal neuron activations on the performance of our method. COINCIDE selects neuron activations from the MSA blocks across the 4th, 8th, 12th, 16th, and 20th layers of the reference model. We experiment with different neuron activations and present the results in Table 6. Transforming the neuron activations from the MSA blocks into boolean vectors by mapping negative values to -1 and positive values to 1 causes a significant performance drop, likely due to substantial information loss, yielding inaccurate clustering and transferability calculation. Extracting neuron activations only from the last layer of the reference model causes a slight performance decrease. As discussed in Section 3.2, LVLM abilities stem from various layers. Hence, relying on the last layer captures only a small portion of these capabilities, leading to the performance decline. Finally, utilizing the neuron activations from the MSA blocks gives superior performance compared to using activations from the FFN blocks. We believe this is because MSA layers use self-attention to share multimodal information, providing richer multimodal understanding.

G The COINCIDE Algorithm

In Algorithm 1, we outline our VIT data selection procedure, which involves several key stages: clustering the data (lines 1-2), calculating the cluster categorical distribution (lines 3-5), and selecting samples from each cluster (lines 6-15).

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 $S_i = \mathbb{E}_j\left(\texttt{cos}(\boldsymbol{e}_i, \boldsymbol{e}_j)\right),\ i \in \{1, 2, \ldots, K\}$ ▷ 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 $\mathcal{C}'_1 \cup \mathcal{C}'_2 \cup \ldots \cup \mathcal{C}'_K$

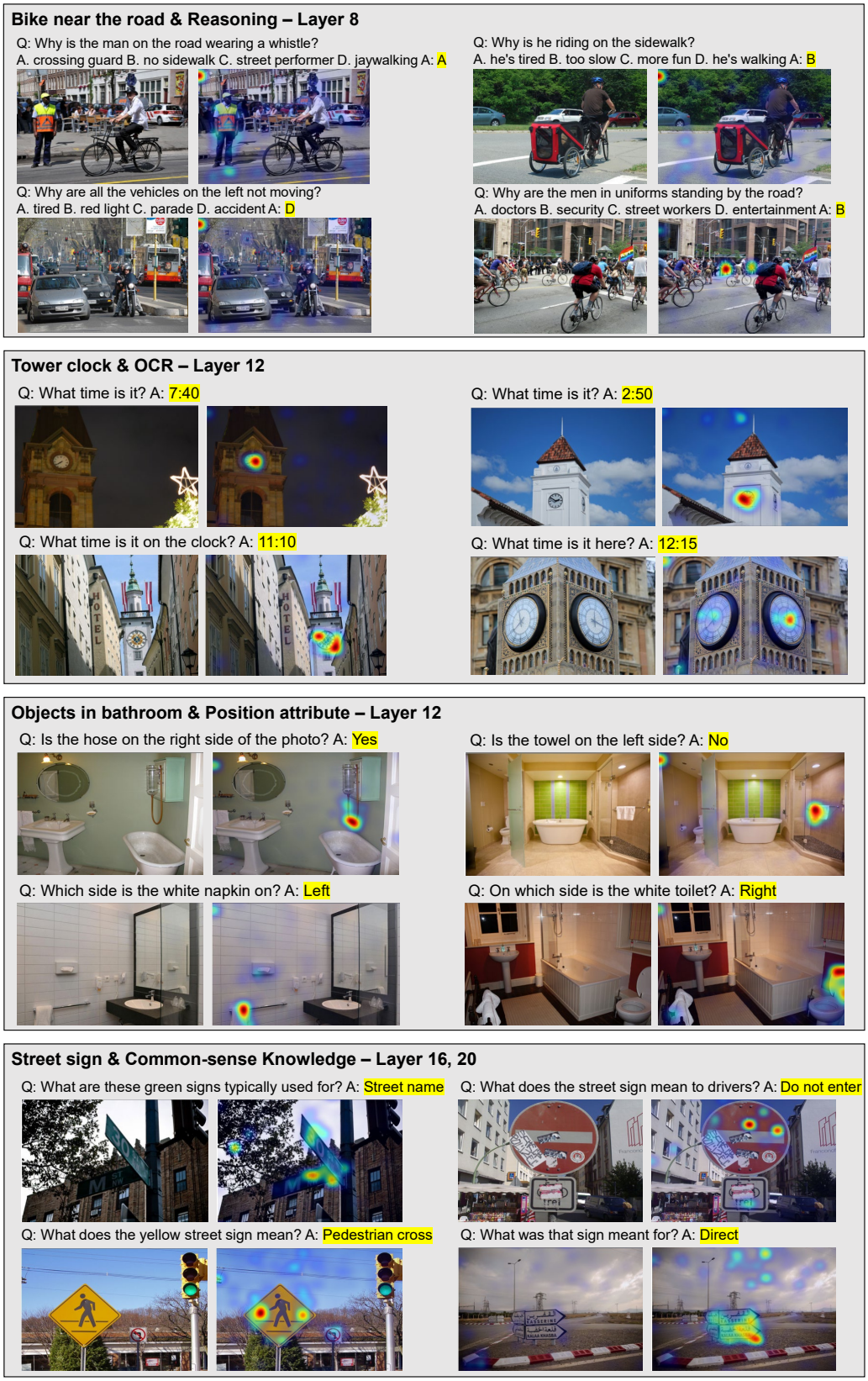


Figure 10: Relevancy maps visualization. We investigate which layer contributes most to the final output of the LVLM. This is done by visualizing relevancy maps of four samples from the same cluster. For each example, the left image is the original, while the right image shows the visualized relevancy map, highlighting regions most relevant to the LVLM output text colored in yellow. The top-left corner of each group explains the VL concept-skill composition and the layer number with the highest relevancy to the output.

Figure 11: Examples of data clusters. We visualize four samples from the same cluster. The top-left corner of each group explains the VL concept-skill composition.

VQAv2 (Books & OCR), Top-1

Q1: What activity must Lynne enjoy doing?
A1: Reading
Q2: What are the objects on?
A2: Table
Q3: Who is the author of the book?
A3: Bill bryson
Q4: What is the name of the book?
A4: Walk in woods
…

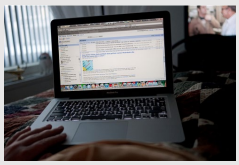
Q1: What color is tintin's dog?
A1: White
Q2: How many books are in the volume?
A2: 8
Q3: Could this be a produce market?
A3: No
Q4: What is the name of the books?
A4: Tintin

Q1: What is the title of the top book?
A1: A place to stand
Q2: Are these library books?
A2: Yes
Q3: Are these books for a college student?
A3: Yes

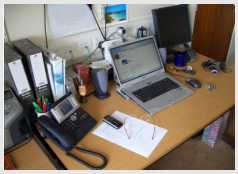
Q1: Is there a photo of a man?
A1: No
Q2: Is there a clock in the picture?
A2: Yes
Q3: Which book was written by John Irving?
A3: Hotel new hampshire

GQA (Electronic devices & Recognition), Top-1

Q1: What kind of furniture is it?
A1: Desk
Q2: Which kind of furniture is that?
A2: Desk
Q3: Are there either any black numbers or words?
A3: No



Q1: Which side of the picture is the TV on?
A1: Right
Q1: Which kind of furniture is this, a sofa or a bed?
A2: Bed
Q3: Is the laptop to the right or to the left of the device on the right?
A3: Left
…



Q1: What piece of furniture are the glasses on?
A1: Desk
Q2: What is the piece of furniture that the glasses are on called?
A2: Desk
Q3: What are the glasses on?
A3: Desk
…

Q1: Which color is the floor?
A1: Gray
Q2: Are the books on the right?
A2: No
Q3: What device is to the left of the lady?
A3: Laptop
Q4: Do you see lamps there?
A4: No
…

A-OKVQA (Objects near human & Color Attribute), Top-1

Q: What colour is the tie on the far right?
A. red
B. yellow
C. orange
D. pink
Answer with the option's letter from the given choices directly.

A: C

Q: What color is the scarf wrapped around the suitcase pulled on the left?
A. red
B. yellow
C. green
D. blue
Answer with the option's letter from the given choices directly.

A: A

Q: The man in the foreground's jacket is the same color as what?
A. banana
B. watermelon
C. orange
D. cherry
Answer with the option's letter from the given choices directly.

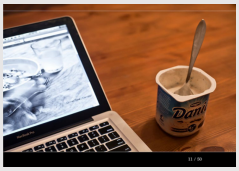
A: A

Q: What color is the neon sign on the second story of this building?
A. pink
B. red
C. violet
D. blue
Answer with the option's letter from the given choices directly.

A: B

LLaVA-Conv (Electronic devices & Conversation + Recognition), Top-1

Q1: What is the woman doing in the image?
A1: The woman in the image is sitting at a table and using a laptop computer.
Q2: Where is the laptop computer placed?
A2: The laptop computer is placed on top of a concrete table.
Q3: Describe the setting where the woman is using her laptop.
…



Q1: What are the two main objects on the wooden desk?
A1: The two main objects on the wooden desk are a laptop and a cup of yogurt.
Q2: Is the yogurt container open or closed?
A2: The yogurt container is open.
Q3: What else is there in the yogurt container?
…

Q1: How many remote controls are in the image?
A`: There are two remote controls in the image.
Q2: How many balls of yarn can be seen in the image?
A2: There is only one ball of yarn in the image.
Q3: What color is the ball of yarn?
…

Q1: What type of game console is in the image?
A1: The image features a Nintendo Wii game console.
Q2: How many controllers are there in the image?
A2: There are two controllers and one nunchuck in the image.
Q3: What type of television is the game
…

Figure 12: High transferability cluster sample visualization. We visualize the samples from the most transferable concept-skill composition for each VL task. The top-left corner of each group explains the VL task type and the VL concept-skill compositions. The VL task type for the group follows the task name where most of the data from the group are associated.

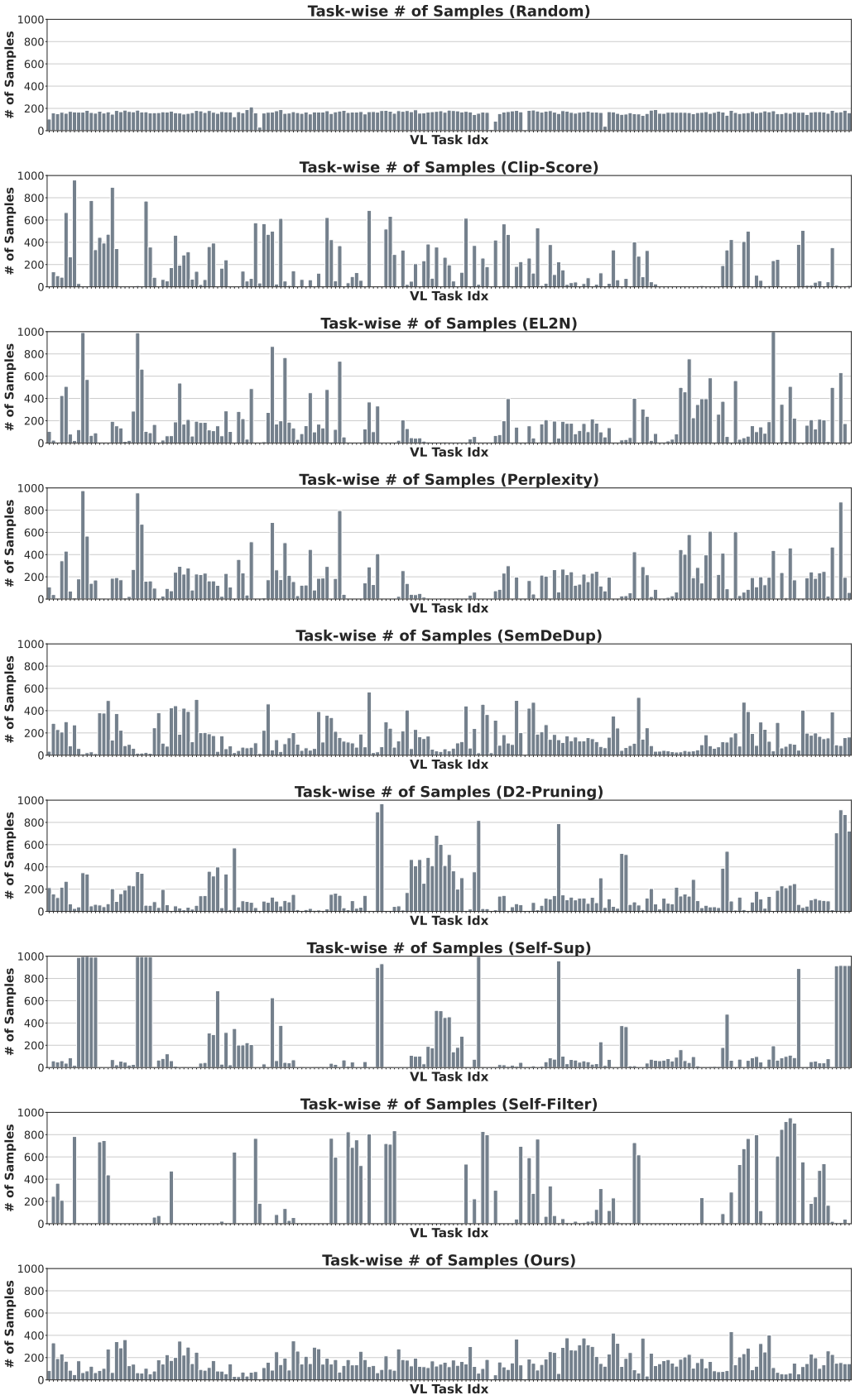


Figure 13: The number of selected samples per VL task in the Vision-Flan VIT dataset. The horizontal axis denotes the VL task index in the dataset, and the vertical axis denotes the number of samples. Baseline methods result in biased coresets. In contrast, our method achieves a more balanced sample selection across diverse tasks, leading to better LVLM generalization.