Capability-Oriented Data Selection

Zicong Zhang [1] **Nanxi Li** [1]

Abstract

Data selection for multimodal large language models (MLLMs) has primarily focused on general-purpose coreset identification, overlooking the capability-specific requirements of downstream tasks. We propose CODS (Capability-Oriented Data Selection), a novel framework that leverages mixture-of-experts (MoE) routing factors alongside gradient and hidden state information to enable targeted data selection based on specific reasoning capabilities. Our approach introduces a capability projecting network that maps multimodal inputs to a structured capability space through a two-stage training paradigm, where representation learning is followed by geometric space adjustment to handle the non-orthogonal nature of multimodal capabilities. Extensive experiments across three diverse benchmarks, LogicOCR, ScienceQA and ViewSpatial for spatial reasoning, demonstrate that CODS consistently outperforms existing data selection methods. Additionally, clustering analysis reveals that our capability-oriented features produce superior cluster quality with higher silhouette scores and Calinski-Harabasz scores (120.1 vs. 97.6) compared to existing methods. These results validate the effectiveness of capability-oriented data selection for enhancing model performance on specialized multimodal reasoning tasks.

1. Introduction

Recently, Multimodal Large Language Models (MLLMs) have demonstrated remarkable performance in various tasks. However, as tasks increasingly demand complex reasoning capabilities across multiple modalities, the need for more sophisticated fine-tuning has emerged (Deng et al., 2025). Drawing from curriculum learning principles (Bengio et al., 2009), an effective data selection method should prioritize capability-specific datapoints that address the model's performance gaps rather than indiscriminately using large, redundant datasets (Wei et al., 2023). This targeted approach not only improves training efficiency but also enables more precise enhancement of specific capabilities.

Despite the clear need for capability-oriented data selection, existing multimodal data selection methods primarily focus on identifying a general-purpose coreset for efficient training. These approaches can be broadly categorized into two groups: Gradient-Based methods (Liu et al., 2024; Wu et al., 2024) and Model-Based methods (Tiong et al., 2024; Lee et al., 2024; Bi et al., 2025). Although they propose different measurements of data importance and achieve promising results, their criteria are not capability-oriented and may not be optimal for specific downstream tasks.

It is natural to ask *How can we design a capability-oriented data selection method that can select data based on the downstream tasks?* and *What is the critical criteria for capability-oriented data selection?* We observe that Mixture-of-Experts (MoE) routing factors, which have been previously underexplored in data selection research, provide valuable insights into modality-specific attention allocation. These routing patterns effectively represent the capabilities required for specific tasks and can serve as a principled basis for capability-oriented data selection.

To harness this insight, we propose a novel **C**apability-**O**riented **D**ata **S**election (**CODS**) framework. Our approach introduces a capability projecting network that systematically maps multimodal inputs to a structured capability space by integrating multiple complementary information sources: gradient-based feature, latent representational states, and crucially, the MoE routing factor distributions. This integration of heterogeneous information streams enables a comprehensive characterization of the capabilities engaged by each training instance. The architecture employs a two-stage optimization process: first projecting data points into a preliminary capability space through supervised contrastive learning, followed by a refinement stage where we implement a geometric repulsion mechanism to address the non-orthogonality of naturally occurring capabilities. To operationalize this framework, we systematically annotate existing multimodal datasets with capability tax-

[*] Equal contribution [1] Zhiyuan College, Shanghai Jiao Tong University. Correspondence to: Nanxi Li <andyc_03@sjtu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

onomies derived from expert knowledge and statistical analysis of model behavior across benchmark tasks. The resulting capability-annotated corpus serves as the foundation for training our projector network, enabling precise capability-oriented data selection that significantly enhances model adaptation efficiency for specialized downstream tasks.

To validate our framework, we conduct comprehensive evaluations across multiple downstream tasks requiring distinct capabilities: LogicOCR (Ye et al., 2025) for logical reasoning on text-rich images, ScienceQA (Lu et al., 2022) for scientific question answering, and ViewSpatial (Li et al., 2025) for spatial reasoning assessment. We compare our capability-oriented data selection method against several baseline approaches, including state-of-the-art methods TIVE (Liu et al., 2024) and COINCIDE (Lee et al., 2024), which represent gradient-based and model-based multimodal data selection paradigms respectively. Following a consistent experimental protocol, we implement Low-Rank Adaptation (LoRA) fine-tuning on the LLaVA-v1.5 architecture (Liu et al., 2023) across all methods to ensure fair comparison. Our experimental results demonstrate that CODS significantly outperforms existing methods, yielding improvements in model performance across all evaluated tasks. Furthermore, we employ quantitative clustering metrics to evaluate the intrinsic quality and discriminative power of our selected data distributions, revealing that our capability-oriented approach achieves superior cluster cohesion and separation compared to alternative methods. These findings not only validate the efficacy of our approach for targeted capability enhancement but also suggest promising broader applications in multimodal data selection for specialized tasks.

To summarize, our contributions are three-fold:

- We first propose to leverage MoE routing factors for capability-oriented data selection, which provides a principled basis for selecting data based on the downstream tasks.
- We introduce CODS, a framework to synthesize diverse information streams for capability projection, which enables more effective data selection oriented for specific capabilities.
- Through LoRA finetuning and clustering experiments, we demonstrate that our proposed framework consistently outperforms existing data selection methods across multiple downstream tasks.

2. Related Work

Multimodal Data Selection. Multimodal data selection methods have recently gained significant attention in the research community. Broadly speaking, existing methods can be categorized into two groups: Gradient-Based methods (Liu et al., 2024; Wu et al., 2024) and Model-Based methods (Tiong et al., 2024; Lee et al., 2024; Bi et al., 2025).

TIVE (Liu et al., 2024) utilizes gradient information to estimate instance influence and task difficulty for core data selection. Similarly, ICONS (Wu et al., 2024) employs a voting mechanism to identify training samples that demonstrate consistent positive impact across different tasks. In the model-based category, COINCIDE (Lee et al., 2024) proposes a coreset selection method by clustering features extracted from multiple layers in a smaller MLLM, while PRISM (Bi et al., 2025) conducts self-pruning selection by leveraging the model's intrinsic token representation.

However, while these existing methods have significantly advanced efficient training through coreset selection, they primarily optimize for general model performance rather than targeting specific capabilities. Though valuable for reducing computational costs, this approach often falls short for specialized downstream tasks requiring particular reasoning capabilities. This highlights the need for a capability-oriented data selection approach that can pave way for effectively supporting complex reasoning tasks involving multiple interrelated capabilities.

Mixture of Vision Experts. Recent research has demonstrated the efficacy of integrating multiple vision models pre-trained on diverse tasks to achieve comprehensive multimodal capabilities. SPHINX (Lin et al., 2023) pioneered this approach by leveraging multiple vision encoders to enhance performance across various tasks. Subsequently, more sophisticated fusion techniques have emerged, including Prismatic (Karamcheti et al., 2024), MOVA (Zong et al., 2024), and EAGLE (Shi et al., 2024). Notably, MOVA (Zong et al., 2024) implements a router-like adapter mechanism whose routing factors provide valuable insights into modality-specific attention allocation across different tasks. Drawing inspiration from these advances in multimodal expertise routing, our work proposes leveraging MoE routing patterns as a principled foundation for capability-oriented data selection, offering a novel approach to identifying task-relevant training samples.

3. Methodology

We propose a novel data-selection framework for Multimodal Large Language Models (MLLMs). Our method maintains a projector to predict task-relevant score affiliating data-selection process. We will introduce the detailed parts of our framework in the following sections.

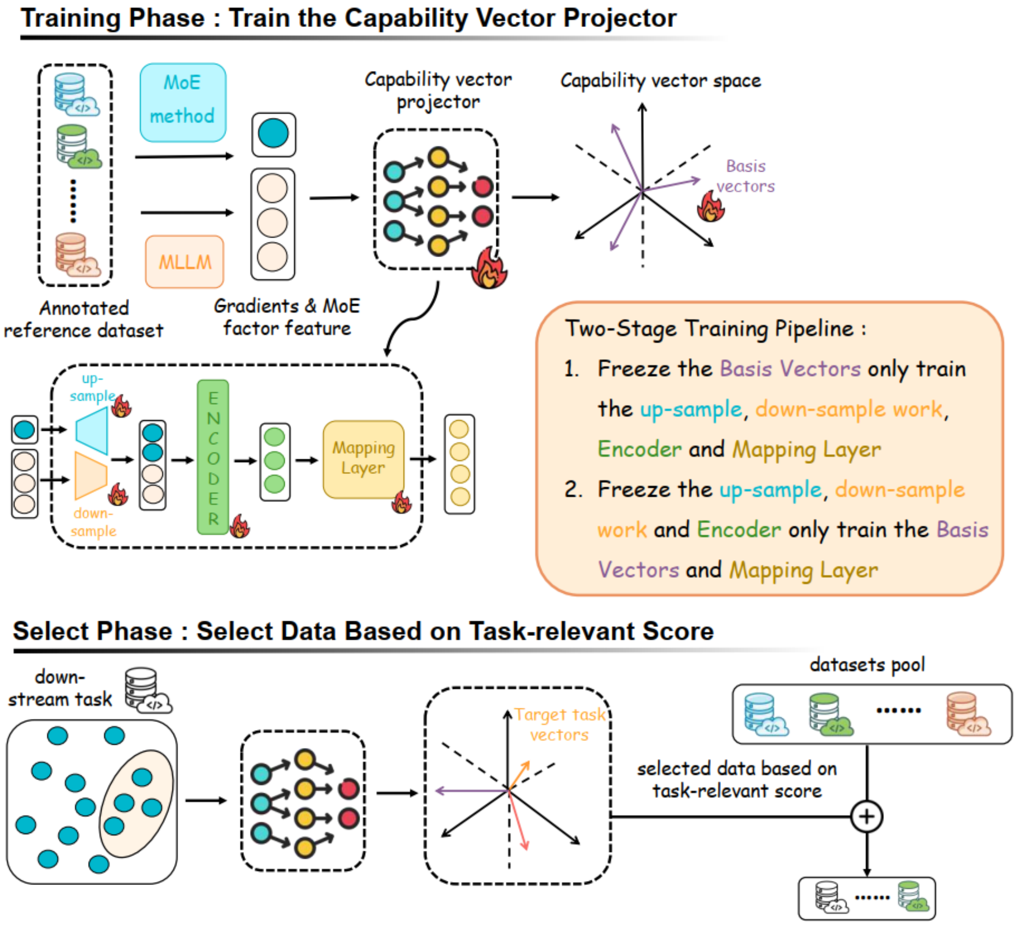
Training Phase : Train the Capability Vector Projector
MoE method
MLLM
Annotated reference dataset
Gradients & MoE factor feature
Capability vector projector
Capability vector space
Basis vectors
up-sample
down-sample
E N C O D E R
Mapping Layer
Two-Stage Training Pipeline :
1. Freeze the Basis Vectors only train the up-sample, down-sample work, Encoder and Mapping Layer
2. Freeze the up-sample, down-sample work and Encoder only train the Basis Vectors and Mapping Layer
Select Phase : Select Data Based on Task-relevant Score
down-stream task
Target task vectors
datasets pool
selected data based on task-relevant score

Figure 1. This is the framework of Capability-Oriented Data Selection (CODS) including training phase and select phase.

3.1. Preliminary

Low Rank Adaptation (LoRA) (Hu et al., 2022) is an effective parameter-efficient tuning method for large language models. With strong versatility, LoRA can be applied to any linear layers. Formally, we denote a linear layer as $h = Wx$ where $x \in \mathbb{R}^{d_{\text{in}}}$ is the input and weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. LoRA adds a low rank decomposed update as :

$$h_{\text{new}} = Wx + \Delta Wx = Wx + \frac{\alpha}{r}BAx$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ are two low rank matrices with r as the chosen rank. Usually, $r \ll \min(d_{\text{in}}, d_{\text{out}})$. And α is the hyper-parameter to control the magnitudes of the changes to the original weight matrix W. On the conventional LoRA tuning diagram, only matrices A and B are updated.

Mixture of Vision Experts. The Mixture of Experts (MoE) architecture represents a significant advancement in multimodal model design, effectively addressing performance limitations and modality-specific biases inherent in single-encoder systems. In MoE-based MLLMs, a collection of K specialized expert encoders operates in parallel, with each expert independently extracting modality-specific features $\{Y_i\}_{i=1}^{K}$ from the input. These representations are then processed by a dynamic gating network that generates expert-specific routing weights $P \in \mathbb{R}^{K}$, which determine the relative contribution of each expert. The weighted aggregation of these diverse representations, formulated as $\hat{X} = \sum_{i=1}^{K} Y_i \cdot P_i$, yields a comprehensive multimodal embedding that captures complementary aspects of the input. This adaptive fusion mechanism, when implemented through multiple sequential blocks, produces increasingly refined representations $\hat{X}_L$ that are ultimately fed to the language model component, providing brilliant guidance for our capability-oriented data selection.

3.2. Foundational Capabilities Mining

Current multimodal benchmarks have designed diverse evaluation tasks that assess different aspects of MLLMs' capabilities, ranging from Scientific Question Answering and Image Captioning to Visual Commonsense Reasoning and

Temporal Reasoning. To effectively support these distinct downstream tasks, we must first identify and differentiate their underlying capability requirements.

In this work, we want to propose leveraging the router weights from mixture of vision encoders as a primary signal to distinguish capability patterns, supplemented by additional information sources including gradients and hidden state representations. By applying unsupervised clustering techniques to these signals and incorporating annotations, we can systematically identify several fundamental capabilities that serve as the basis for our data selection framework. This capability taxonomy enables targeted data selection that aligns with the specific requirements of downstream tasks rather than relying on general-purpose coreset selection.

After systematizing the foundational capabilities, we collected abundant open-source multimodal datasets and annotated them with the relevant foundational capabilities label. These labels makes difference in the projector training process that we will discussed in the Section 3.3.

3.3. Capability Projecting Network

The core of our approach is a capability projector that maps data samples to a capability space where each dimension represents the relevance to a specific foundational capability.

Let $\mathcal{X}$ denote the input space of multimodal samples and $\mathbb{R}^f$ represent the embedding space of our MLLM. Our capability projector is defined as a function $h_\theta : \mathbb{R}^f \rightarrow \mathbb{R}^d$ that maps model features to a d-dimensional capability vector:

$$\mathbf{c} = h_\theta(f_{\text{MLLM}}(x)) \tag{1}$$

where $f_{\text{MLLM}}(x)$ represents the model features (including MoE routing factors, hidden states, and gradients) for input $x \in \mathcal{X}$, and $\mathbf{c} \in \mathbb{R}^d$ is the capability vector with each dimension c_i representing the relevance to capability i.

As explained in framework1, the projector architecture is designed to enhance the task-specific data selection. First, model features (including MoE routing factors, hidden states, and gradients) undergo adaptive resizing via parallel up-sample and down-sample modules to balance the feature dimensions. Conditioned features are then processed by a encoder module to capture latent relationships. Finally, sequential mapping layers transform encoded representations into the capability vector space.

3.4. Two-Stage Training Paradigm

To realize better training performance of the projecting network, we design a novel two-stage training paradigm, which can effectively solve optimization problems in multi-modal capability projection. This training pipeline decouples the ability space learning into two stages: (1) Representation Learning and (2) Space Adjustment, effectively balancing discriminative power for multi-label classification with task relevance.

3.4.1. Representation Learning

For Representation Learning stage, the primary goal is to train the essential capability projector to learn basic feature representation and classification skills. We initialize the basis vectors to be orthogonal and freeze them, while only training the encoder and related components. Given that input samples typically exhibit multi-label annotations, indicating simultaneous relevance to multiple capabilities and suffer from significant class imbalance, we adopt Focal Loss (Lin et al., 2017) as our primary objective:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{focal}} = \frac{1}{N \times C} \sum_{i=1}^{N} \sum_{c=1}^{C} \alpha \cdot (1 - p_t^{(i,c)})^\gamma \cdot \mathcal{L}_{\text{BCE}}^{(i,c)} \tag{2}$$

Here, N denotes batch size, C the number of foundational capabilities, and $p_t^{(i,c)}$ the model's calibrated probability for sample i belonging to capability c. The hyperparameters α (set to 0.25) and γ (set to 2.0) systematically mitigate class imbalance by reducing the contribution of well-classified examples while amplifying gradients from challenging cases. The term $\mathcal{L}_{\text{BCE}}$ represents the Binary Cross-Entropy component.

We train this stage for 20 epochs with a learning rate of 2×10^{-3}, establishing robust feature representations while maintaining a stable orthogonal basis vectors for subsequent refinement.

3.4.2. Space Adjustment

Real-world capabilities exhibit complex interdependencies, for instance, Spatial Reasoning inherently builds upon Object Localization competence. So it's crucial to optimize the geometric structure of the embedding space, make the similarity between basis vectors closer to the real foundational capability correlation and modify the layout of projected vectors. So in this stage, we freeze the encoder and related components, only training the sequential mapping layers and basis vectors to adjust the geometric structure and projecting performance.

To ensure the same kind of sample is closer in the capability space, and the different kind of sample is farther away, we add the Contrastive Loss to modify the layout of projected vectors.

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{P}_i} \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)} \tag{3}$$

where $\mathbf{z}_i$ is the L2-normalized feature vector and $\mathcal{P}_i$ represents the set of positive samples for anchor i (samples sharing at least one label with i)

In order to ensure that the basis vectors are as dispersed and independent as possible during the training process, we add the Repulsion Loss.

$$\mathcal{L}_{\text{repulsion}} = \frac{1}{C(C-1)} \sum_{i=1}^{C} \sum_{j \neq i} \max\left(|\langle \mathbf{b}_i, \mathbf{b}_j \rangle| - \delta, 0\right) \tag{4}$$

where $\mathbf{b}_i$ indicates the L2-normalized basis vector corresponding to i-th foundational capability, $\langle \mathbf{b}_i, \mathbf{b}_j \rangle$ is the cosine similarity between two basis vectors and δ is a margin to add some tolerance.

Meanwhile, to maintain the discriminative classification skills, we also keep the Focal Loss. So the total loss during the Space Adjustment stage combines three complementary components :

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{focal}} + \lambda_c \cdot \mathcal{L}_{\text{contrastive}} + \lambda_r \cdot \mathcal{L}_{\text{repulsion}} \tag{5}$$

In this stage, we set the coefficient $\lambda_c = 0.1, \lambda_r = 0.05$, the margin $\delta = 0.1$ and train the model for 6 epochs with learning rate as 5×10^{-4}

4. Evaluation Experiment

4.1. Settings

To evaluate the efficacy of our proposed method, we conduct extensive experiments across diverse multimodal benchmarks that assess distinct capabilities. Specifically, we select LogicOCR (Ye et al., 2025), which evaluates complex logical reasoning on text-rich images; ScienceQA (Lu et al., 2022), which encompasses a broad spectrum of scientific topics in image-text question pairs; and ViewSpatial (Li et al., 2025), which examines spatial localization recognition capabilities in visual language models. Our experimental protocol involves providing 100 exemplars as reference instances for each capability category. Subsequently, we employ clustering algorithms to select appropriate training samples using each data selection methodology. For the training data pool, we utilize a curated subset of the LLaVA-665K dataset introduced by Liu et al. (2023), ensuring consistent comparison across all methods.

4.2. Baselines

We evaluate our method against several baseline approaches. The baselines include: (1) *Repeat*, which simply repeats the given reference examples; (2) *Random*, which randomly samples from the data pool; and (3) existing data selection methods for fair comparison, including TIVE (Liu et al., 2024) and COINCIDE (Lee et al., 2024). For the TIVE, COINCIDE and our proposed CODS, we adopt a consistent evaluation protocol: we first extract features using each method's respective feature generation approach, then perform clustering on these features using the reference data, and finally select data points with minimum distance to the identified cluster centers.

To be precise, we value each datapoint x in the data pool with a score $s(x)$, which is computed as follows:

$$s(x) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \min_{x' \in \mathcal{D}_c} d(f(x), f(x')) \tag{6}$$

where $\mathcal{C}$ is the set of reference data, $f(x)$ is the feature representation of x, and $d(\cdot, \cdot)$ is a distance metric (e.g., cosine distance). The selected data points are those with the lowest scores.

4.3. Evaluation Experiment Details

We conducted experiments using LoRA fine-tuning on the LLaVA-v1.5 model. For all methods, we maintained consistent hyperparameters: LoRA rank $r = 8$, scaling factor $\alpha = 32$, and learning rate 1×10^{-5}. Each model was trained for three epochs with the dataset partitioned into 80% training and 20% validation sets. All experiments were performed on a single NVIDIA RTX 4090 GPU with a batch size of 4 to ensure fair comparison across methods.

4.4. Main Results

As shown in Table 1, our proposed CODS framework consistently outperforms all baseline methods across the three evaluated benchmarks. On LogicOCR, CODS achieves a 10.9% absolute improvement over the baseline LLaVA-v1.5 model, demonstrating its effectiveness in enhancing complex logical reasoning on text-rich images. For ScienceQA, our method delivers a substantial 16.3% performance gain, indicating its strong capability in improving scientific reasoning across diverse topics.

The most significant improvement is observed on ViewSpatial, where CODS achieves a remarkable 13.2% absolute gain over the baseline and substantially outperforms all competing methods by a large margin (11.6% higher than the second-best method). This pronounced advantage suggests that our capability-oriented approach is particularly effective for tasks requiring spatial reasoning, where modality-specific routing patterns likely provide especially valuable signals for identifying relevant training examples.

While the improvements on LogicOCR and ScienceQA are somewhat marginal compared to strong baselines like Repeat (0.3% and 1.1% respectively), we speculate that this is

Method	LogicOCR	ScienceQA	ViewSpatial
LLaVA-v1.5	26.4	48.2	37.1
+ Repeat	<u>37.0</u>	<u>63.4</u>	38.6
+ Random	33.9	61.1	<u>38.7</u>
+ TIVE	36.9	63.2	38.0
+ COINCIDE	36.7	63.2	38.0
+ CODS (Ours)	**37.3** (+10.9)	**64.5** (+16.3)	**50.3** (+13.2)

Table 1. Accuracy on different data selection methods on LogicOCR, ScienceQA and ViewSpatial. **Bold** indicates the best result in each column, <u>underlined</u> indicates the second-best result. Green deltas show improvements over the baseline LLaVA-v1.5.

primarily due to limitations in the training data pool. Modern complex reasoning benchmarks like these often require diverse, high-quality examples that might not be sufficiently represented in the current LLaVA-665K subset. This hypothesis is supported by the relatively small performance gap between different selection methods on these benchmarks, suggesting that the data pool itself may be the limiting factor rather than the selection methodology.

The consistent superiority of our approach across all benchmarks, despite these data constraints, validates the fundamental premise of capability-oriented data selection. We leave the expansion of the data pool and more comprehensive validation of our method across additional benchmarks and more models for future work, which we anticipate will further amplify the performance advantages of our capability-oriented selection approach.

5. Visualization

Since the geometry of embeddings significantly impacts clustering performance and downstream data selection quality, we visualize the embedding features extracted from a curated subset of the LLaVA-665K dataset using t-SNE. We compare our CODS method with TIVE (Liu et al., 2024) and COINCIDE (Lee et al., 2024), as illustrated in Figure 2.

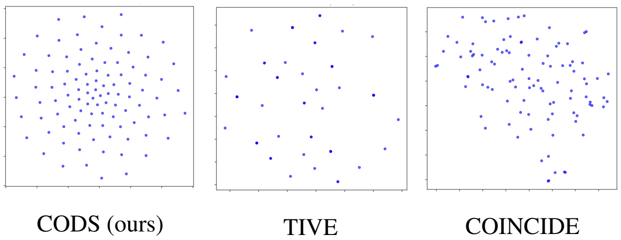

Figure 2. t-SNE visualization of features extracted from curated subset of the LLaVA-665K dataset using different methods

To further quantify embedding geometry, we apply K-means clustering and adopt two widely-used evaluation metrics, the Silhouette Score and Calinski-Harabasz Score. The Silhouette Score asses the quality of internal clustering by taking into account both cohesion and separation. The Calinski - Harabasz Score is another evaluation metric that measures the compactness of clusters by the ratio of the between-cluster dispersion to the within-cluster dispersion. Higher Silhouette Score and C-H Score indicate better separation and cohesion. As shown in Table 2, our proposed CODS framework consistently outperforms other baseline methods across these evaluation metrics, demonstrating its efficacy in learning robust representations with improved embedding-space geometry. These results further corroborate CODS's exceptional data selection performance.

Method	Sil. Score	CH Score
TIVE	0.310	62.3
COINCIDE	0.388	97.6
CODS (Ours)	**0.445**	**120.1**

Table 2. Clustering quality comparison across different data selection methods. Higher scores indicate better cluster separation and cohesion. Sil. Score = Silhouette Score, CH Score = Calinski-Harabasz Score.

6. Conclusion and Discussion

In this work, we introduced CODS (Capability-Oriented Data Selection), a novel framework that leverages mixture-of-experts routing factors alongside gradient and hidden state information to enable targeted data selection based on specific reasoning capabilities for multimodal large language models. Our approach addresses the limitation of existing data selection methods that focus on general-purpose coreset identification rather than capability-specific requirements. Through a capability projecting network and two-stage training paradigm, CODS systematically maps multimodal inputs to a structured capability space, enabling precise identification of task-relevant training samples. Extensive experiments across LogicOCR, ScienceQA, and ViewSpatial demonstrate consistent improvements. Additionally, clustering analysis reveals superior feature quality with higher silhouette and Calinski-Harabasz scores compared to existing methods.

Despite promising results, our evaluation is constrained by the current training data pool size and diversity. Nevertheless, our clustering results demonstrate superior feature quality with significantly higher silhouette and Calinski-Harabasz scores compared to existing methods, suggesting broader applications for CODS in selecting data for complex reasoning tasks. Future directions include leveraging our framework for efficient curation of high-quality reasoning datasets, developing dynamic capability mapping techniques, and extending the approach to other modalities. As models advance toward more sophisticated reasoning capabilities, our principled data selection approach offers a valuable foundation for targeted performance enhancement in specialized multimodal reasoning tasks.

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Bi, J., Wang, Y., Yan, D., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

Deng, H., Zou, D., Ma, R., Luo, H., Cao, Y., and Kang, Y. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., and Sadigh, D. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

Lee, J., Li, B., and Hwang, S. J. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*, 2024.

Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., Lu, W., and Zhuang, Y. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025. URL https://arxiv.org/abs/2505.21500.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP (99):2999–3007, 2017.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2023. URL https://api.semanticscholar.org/CorpusID:263672058.

Liu, Z., Zhou, K., Zhao, W. X., Gao, D., Li, Y., and Wen, J.-R. Less is more: High-value data selection for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.-A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

Tiong, A. M. H., Zhao, J., Li, B., Li, J., Hoi, S. C., and Xiong, C. What are we measuring when we evaluate large vision-language models? an analysis of latent factors and biases. *arXiv preprint arXiv:2404.02415*, 2024.

Wei, L., Jiang, Z., Huang, W., and Sun, L. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*, 2023.

Wu, X., Xia, M., Shao, R., Deng, Z., Koh, P. W., and Russakovsky, O. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024.

Ye, M., Zhang, J., Liu, J., Du, B., and Tao, D. Logicocr: Do your large multimodal models excel at logical reasoning on text-rich images? *arXiv preprint arXiv:2505.12307*, 2025.

Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., and Liu, Y. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.