ICML
International Conference
On Machine Learning

# Foundation Model Insights and a Multi-Model Approach for Superior Fine-Grained One-shot Subset Selection



Zhijing Wan


Zhixiang Wang

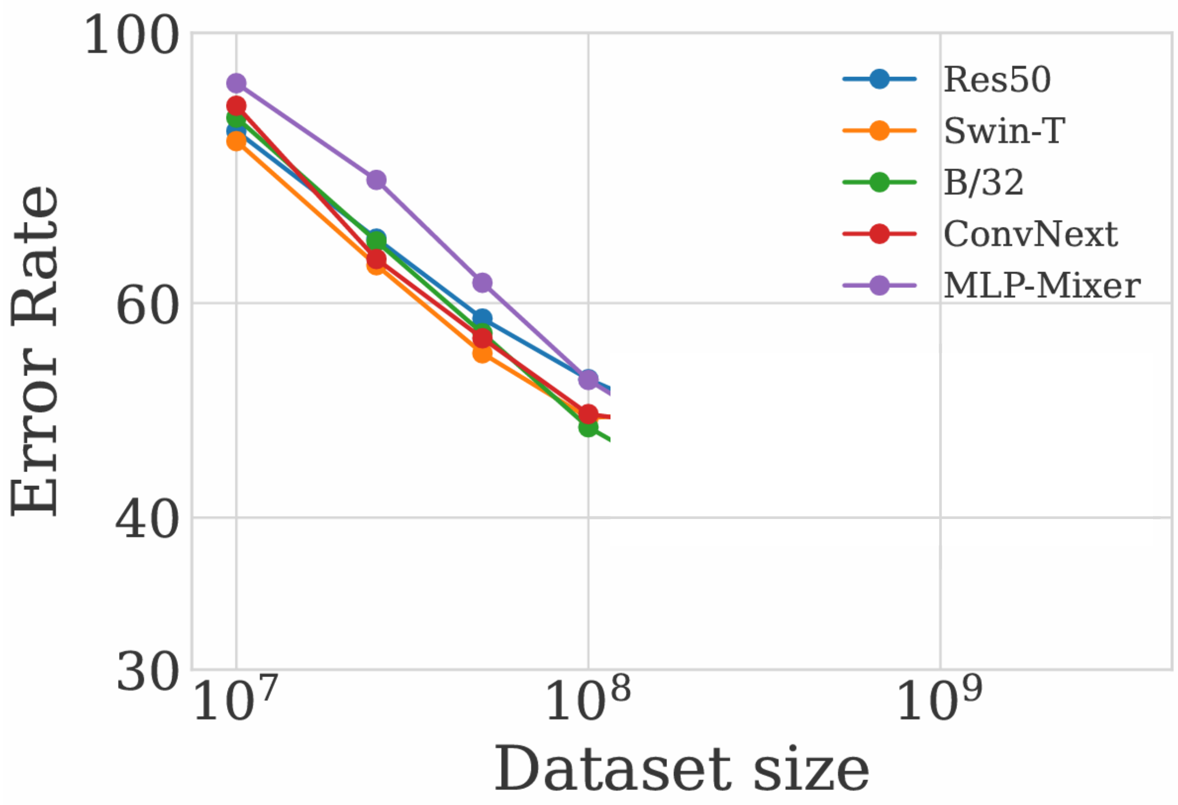

Zheng Wang


Xin Xu


Shin'ichi Satoh


July 16, 2025

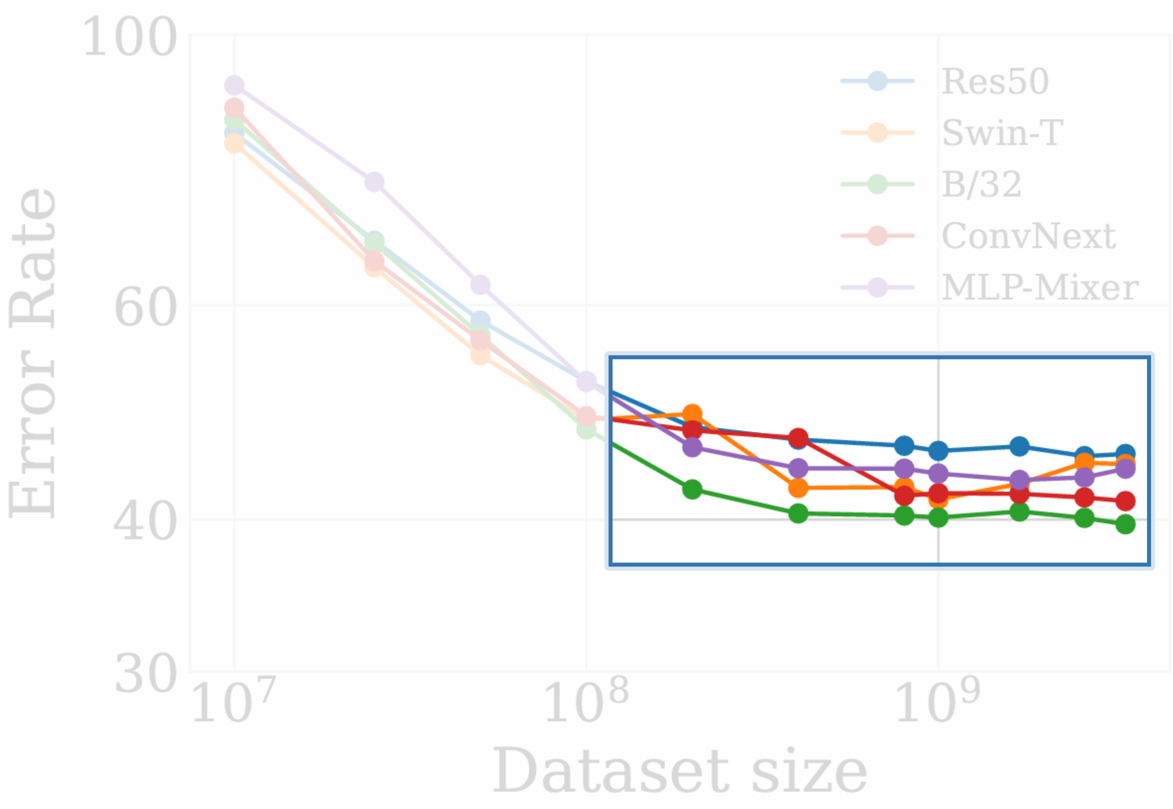
# Data Explosion Fuels Deep Learning




Dataset size V.S. ImageNet Error Rate [1]

[1] Li Z, et al. Scaling (down) clip: A comprehensive analysis of data, architecture, and training strategies. arXiv preprint arXiv:2404.08197, 2024.

# More Data ≠ Better




Dataset size V.S. ImageNet Error Rate [1]

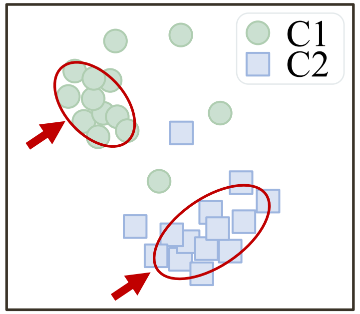
[1] Li Z, et al. Scaling (down) clip: A comprehensive analysis of data, architecture, and training strategies. arXiv preprint arXiv:2404.08197, 2024.

# More Data ≠ Better

*Storage* *Computation* *Annotation*

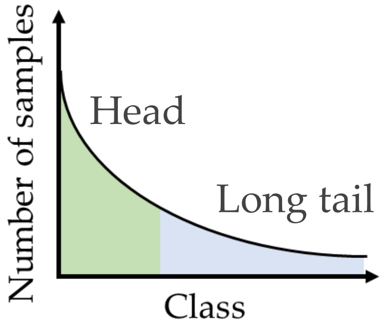
Cost

*Redundancy*




*Label noise*


*Class imbalance*

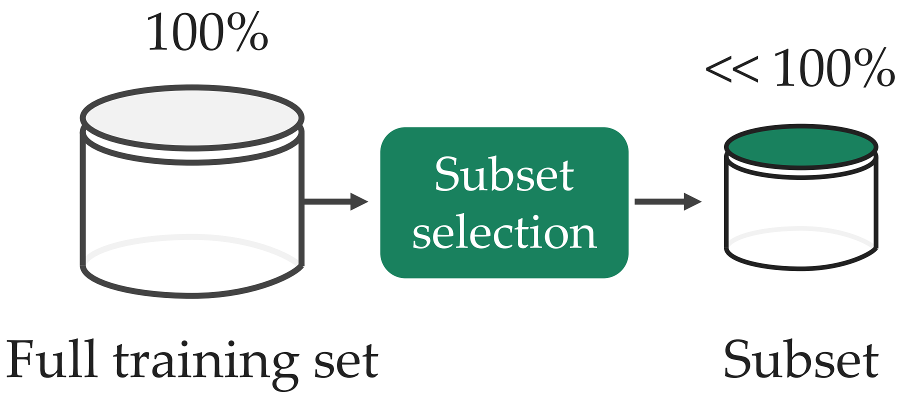



Data quality

Wan Z, et al. A survey of dataset refinement for problems in computer vision datasets, CSUR2024.

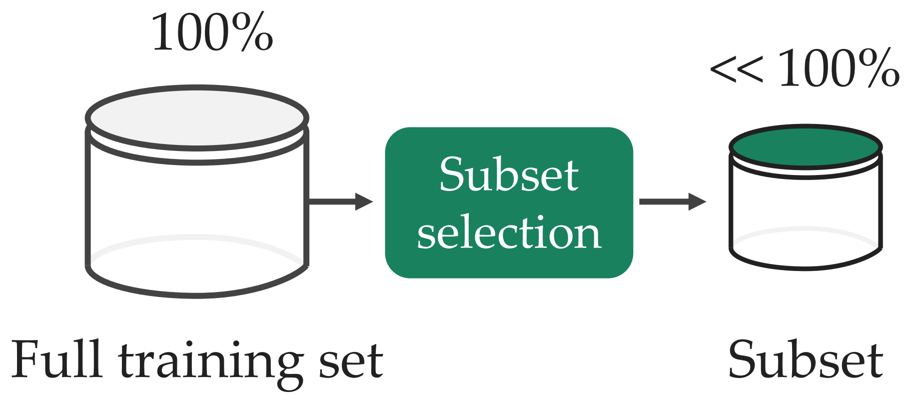
# **Subset Selection:** Balancing Data Volume and Quality

**Goal**: Identify the most informative samples to enable *efficient* training without significantly compromising model performance.

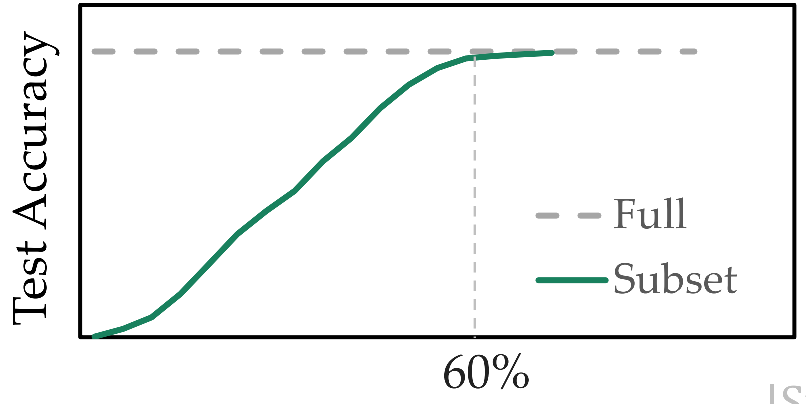


# **Subset Selection**: Balancing Data Volume and Quality

**Goal**: Identify the most informative samples to enable *efficient* training without significantly compromising model performance.



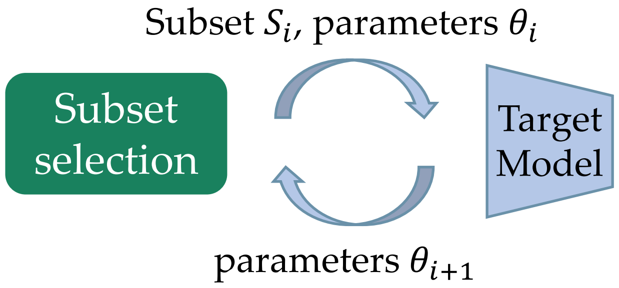

60% selected data yields comparable performance to full-data training on CIFAR-10 [1]




$$\text{Sampling rates} = \frac{|\text{Subset}|}{|\text{Full training set}|}$$

[1] Yang S, et al. Dataset pruning: Reducing training data by examining generalization influence. ICLR2023.

# **Subset Selection**: Two Main Paradigms




### (a) Adaptive Subset Selection

[Karanam et al., 2022; Killamsetty et al., 2022]

- **Subset Selection:**

$$S_i = \mathcal{S}(\theta_i),$$

where $\mathcal{S}(\theta_i)$ is a subset selection function depending on current model parameters.

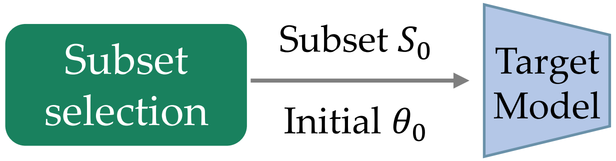
- **Target Model Update:**

$$\theta_{i+1} = \theta_i - \eta \nabla_\theta f(\theta_i; S_i)$$

- **Feedback Loop:**

$$\theta_0 \rightarrow S_0 \rightarrow \ldots \rightarrow \theta_i \rightarrow S_i \rightarrow \theta_{i+1} \rightarrow S_{i+1} \rightarrow \ldots$$

until the target model training converges.




### (b) One-shot Subset Selection

[Xia et al., 2024; Yang et al., 2024]

- **Subset Selection:**

$$S_0 = \mathcal{S}(\theta_{\text{pre-trained}}),$$

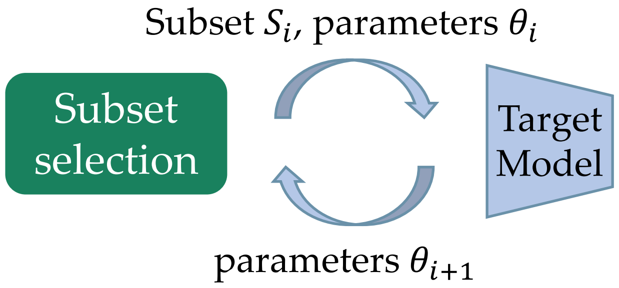
where $\mathcal{S}(\theta_{\text{pre-trained}})$ is a subset selection function based on a pre-trained model parameters.

- **Target Model Training:**

$$\theta^* = \arg\min_\theta f(\theta; S_0)$$

- **No feedback loop**

# **Subset Selection**: Two Main Paradigms

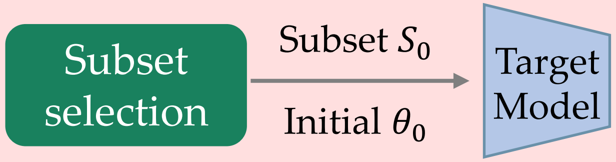



**(a) Adaptive Subset Selection**

[Karanam et al., 2022; Killamsetty et al., 2022]

Iterative selection

- ✗ High selection cost and time-consuming
- ✗ Requires full-dataset access




**(b) One-shot Subset Selection**

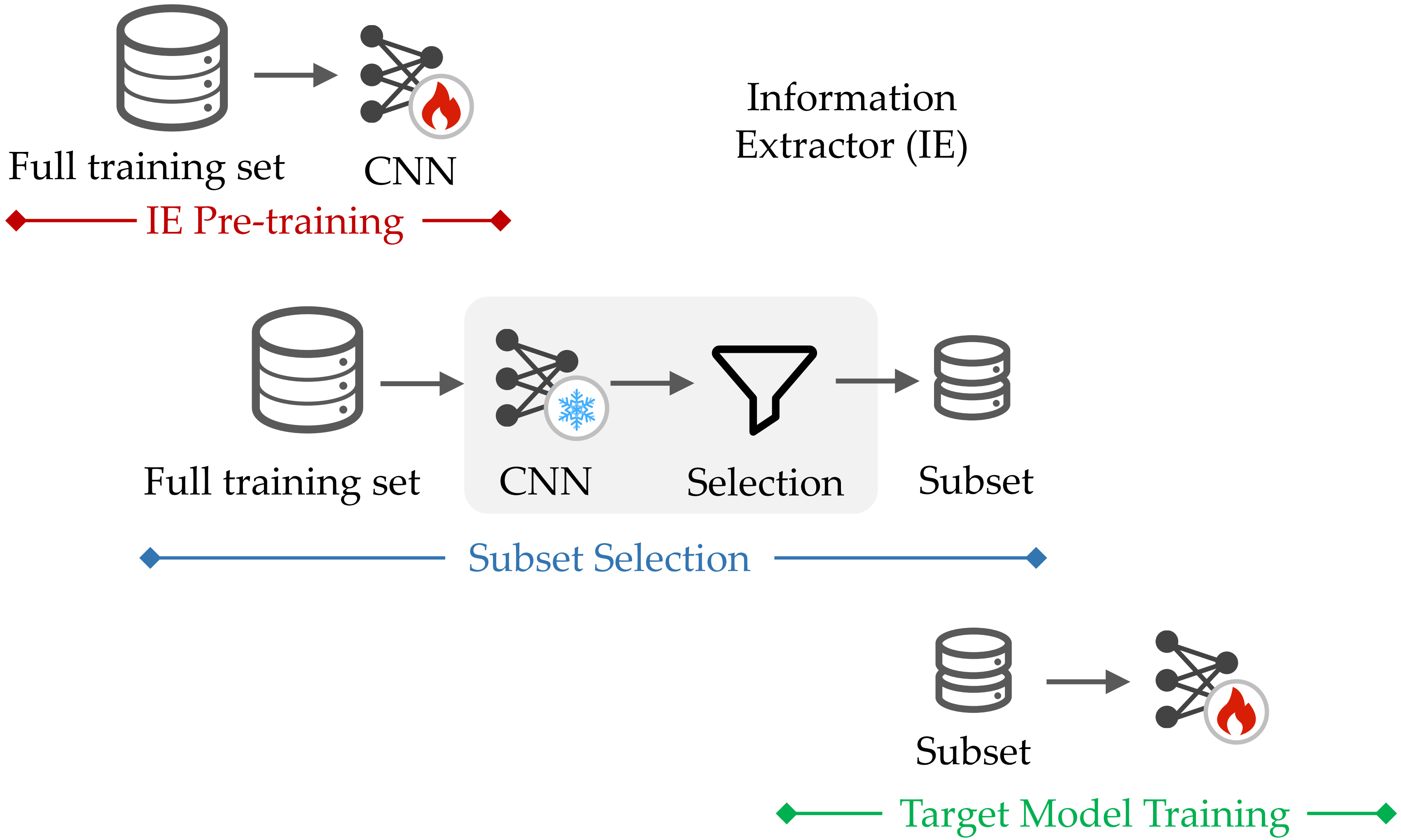
[Xia et al., 2024; Yang et al., 2024]

Single-pass

- ✓ Highly efficient and scalable
- ✓ No full-set storage after selection

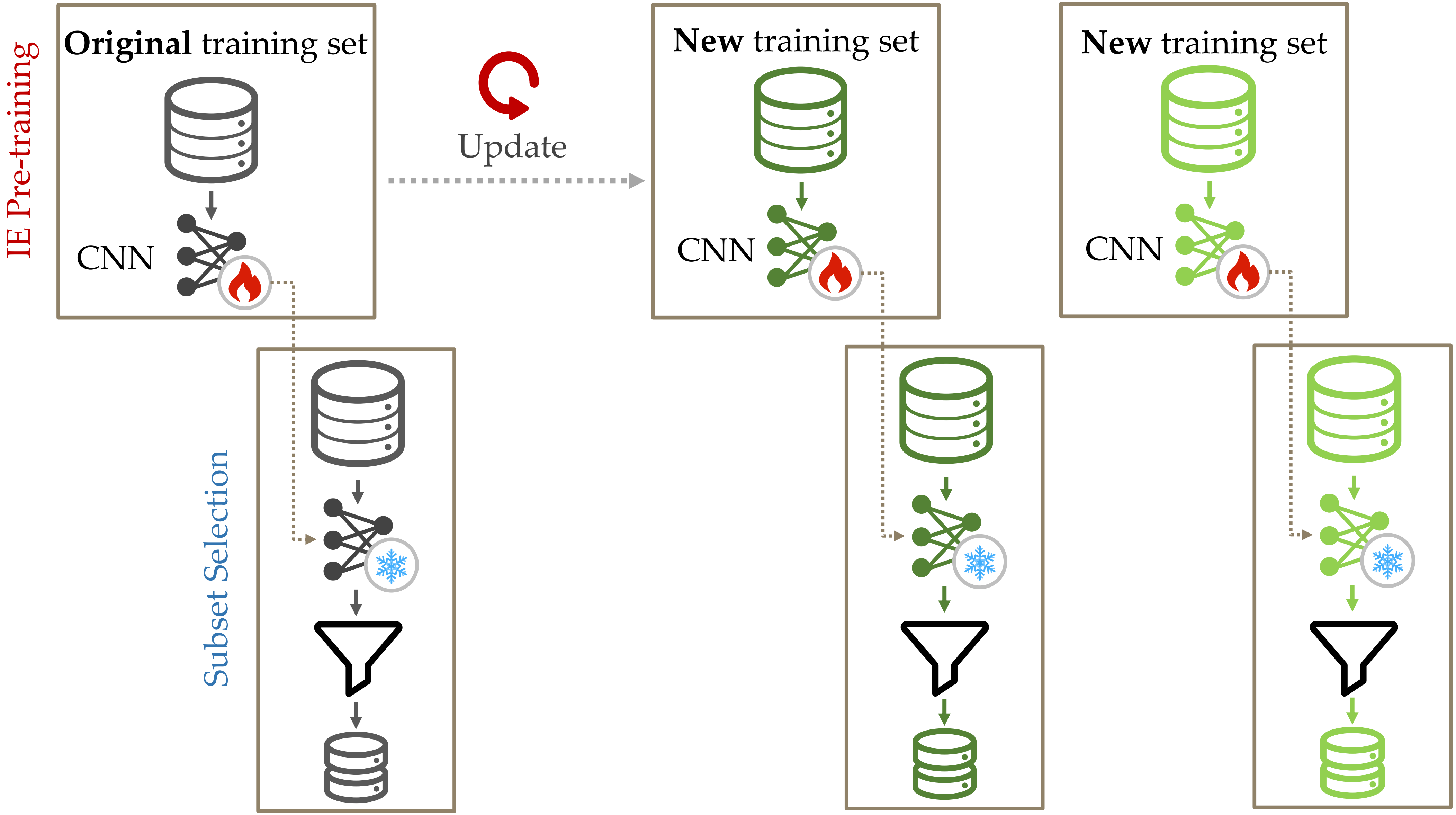
Our focus

# **One-shot Subset Selection:** Existing Pipeline and Challenge



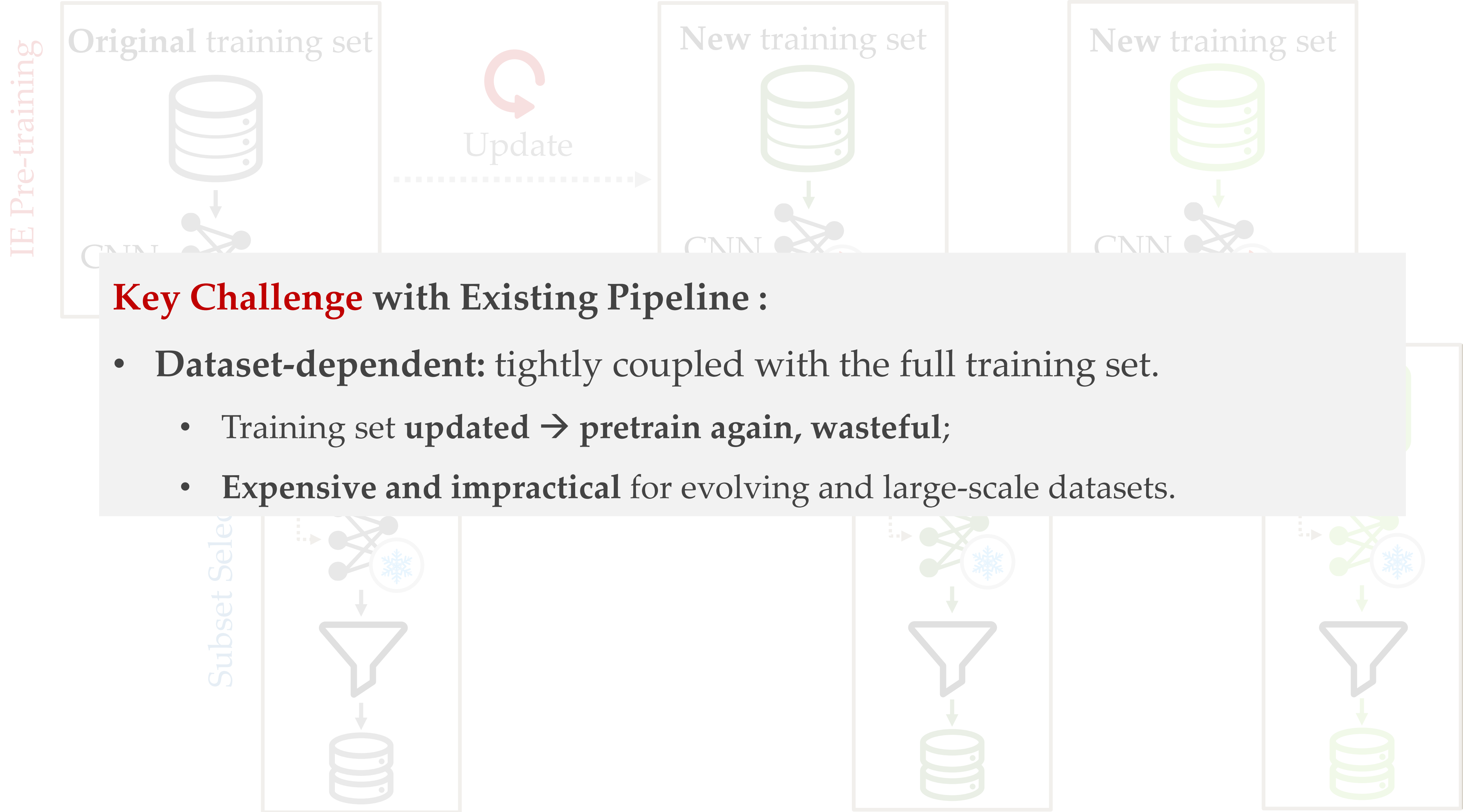

Yang, S., et al. Mind the boundary: Coreset selection via reconstructing the decision boundary. ICML2024.
Zhang, X., et al. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. CVPR 2024.

# **One-shot Subset Selection:** Existing Pipeline and Challenge




Yang, S., et al. Mind the boundary: Coreset selection via reconstructing the decision boundary. ICML2024.
Zhang, X., et al. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. CVPR 2024.

# **One-shot Subset Selection:** Existing Pipeline and Challenge

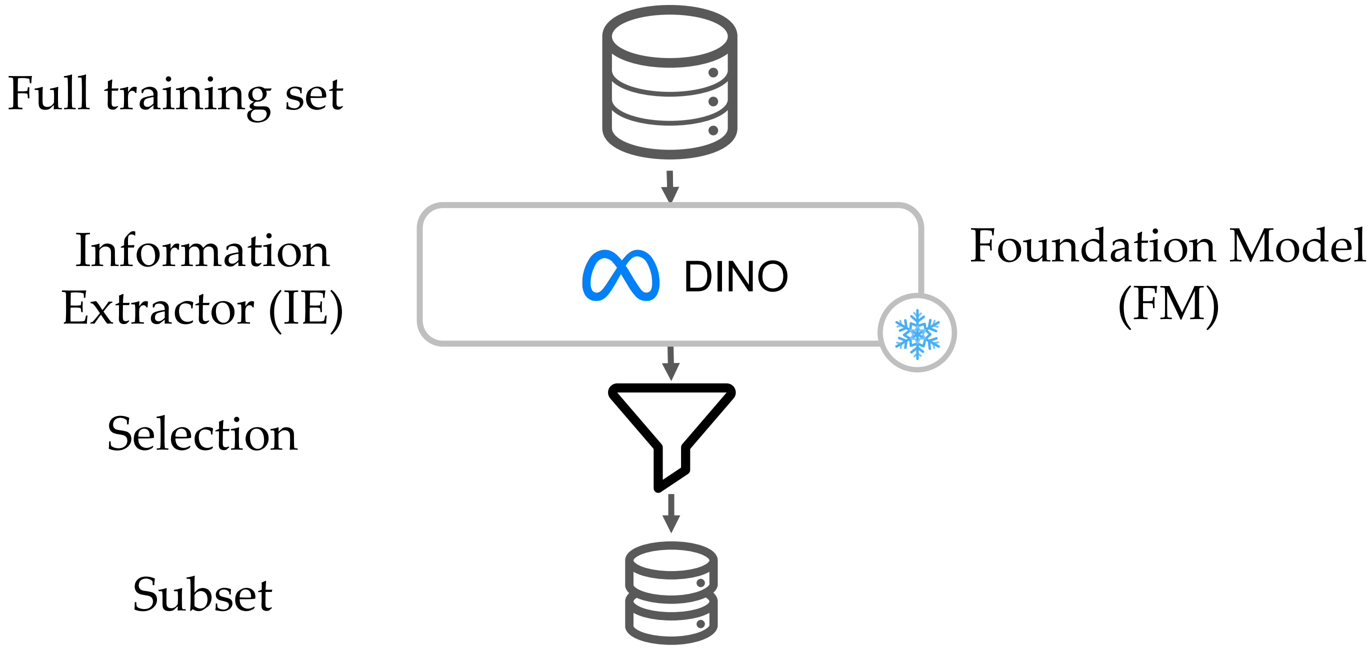



**Key Challenge with Existing Pipeline :**

- **Dataset-dependent:** tightly coupled with the full training set.
  - Training set **updated → pretrain again, wasteful**;
  - **Expensive and impractical** for evolving and large-scale datasets.

Yang, S., et al. Mind the boundary: Coreset selection via reconstructing the decision boundary. ICML2024.
Zhang, X., et al. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. CVPR 2024.

# FM-based Subset Selection: A Dataset-Agnostic Alternative




Xie, Y., et al. Towards free data selection with general-purpose models. NeurIPS 2023.
Killamsetty, K., Milo: Model-agnostic subset selection framework for efficient model training and tuning. Arxiv 2023.

# **FM-based Subset Selection:** A Dataset-Agnostic Alternative

**Key Advantage of Foundation Models:**

- **Strong generalization** across domains and distributions.
  - **No task-specific pretraining** required;
  - **Eliminate dataset dependency** in subset selection;
  - **Scalable & practical** for large, diverse, or evolving data.

Xie, Y., et al. Towards free data selection with general-purpose models. NeurIPS 2023.
Killamsetty, K., Milo: Model-agnostic subset selection framework for efficient model training and tuning. Arxiv 2023.

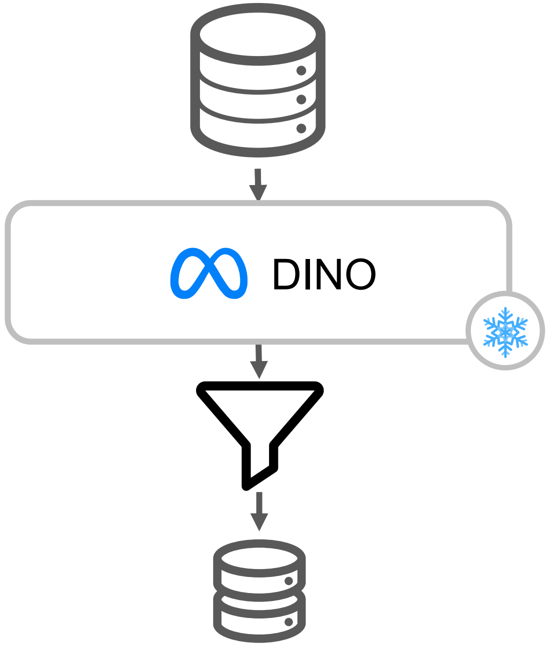
# **FM-based Subset Selection:** A Dataset-Agnostic Alternative

Full training set

Information
Extractor (IE)

Selection

Subset




Foundation Model
(FM)

Zero re-training cost

Xie, Y., et al. Towards free data selection with general-purpose models. NeurIPS 2023.
Killamsetty, K., Milo: Model-agnostic subset selection framework for efficient model training and tuning. Arxiv 2023.

# **FM-based Subset Selection:** Limitations

Full traini

Inform
Extract

Selec

Sub

**Problems with Existing FM-based Subset Selection:**

| **Existing Research** | **Real-World Challenges** |
|---|---|
| IE: A **single** FM (i.e., DINO) | A **spectrum** of FMs |
| **Perfect** task datasets:<br>• Mainly coarse-grained<br>• Clean labels<br>• Class balance | **Not perfect** task datasets:<br>• Fine-grained<br>• Noisy labels<br>• Class imbalance |

Xie, Y., et al. Towards free data selection with general-purpose models. NeurIPS 2023 .
Killamsetty, K., Milo: Model-agnostic subset selection framework for efficient model training and tuning. Arxiv 2023.

# **FM-based Subset Selection:** A Dataset-Agnostic Alternative


## Question

Can FM-based subset selection truly outperform traditional IE-based methods across diverse datasets?

# **Single-Model Study**: Setting*

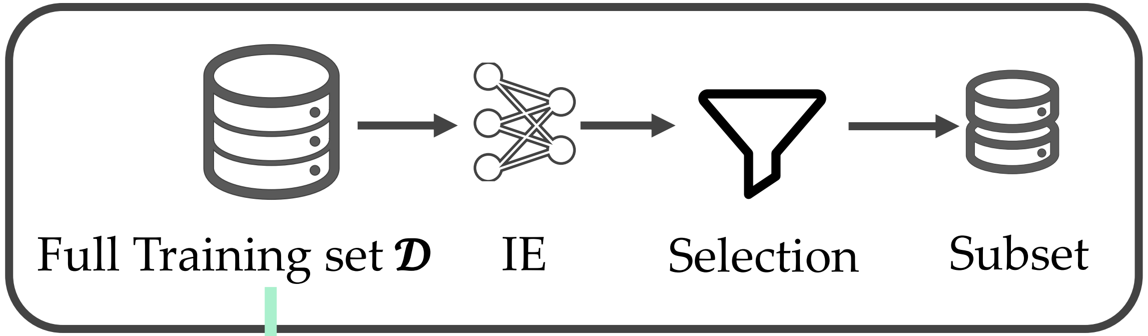
*See paper for details

- **5** datasets × **3** types of IEs × **4** selection methods × **3** sampling rates

# **Single-Model Study**: Setting*

*See paper for details

- **5** datasets × **3** types of IEs × **4** selection methods × **3** sampling rates

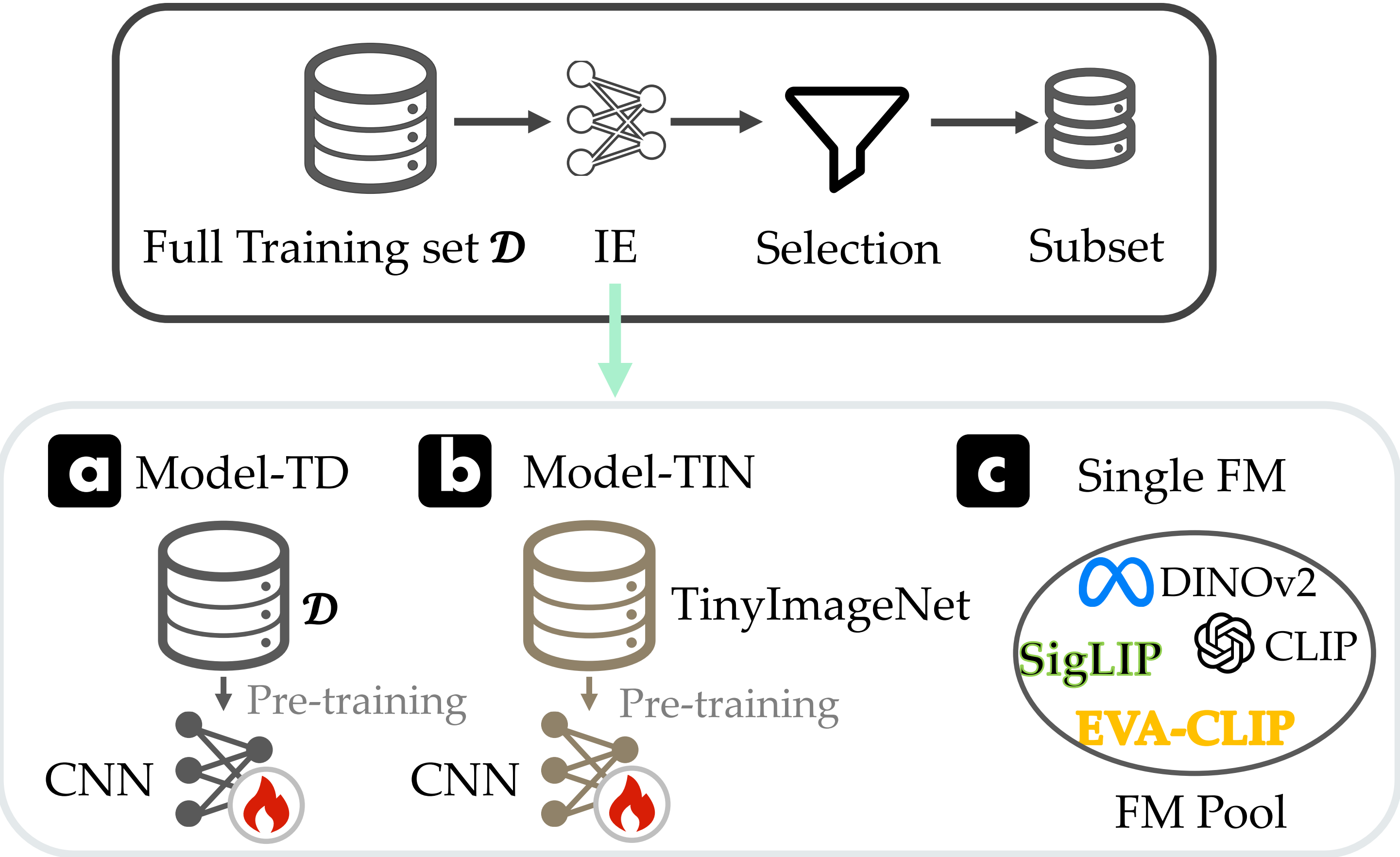



| **Diverse Datasets*** | Grained Level | Noise label | Class Imbalance |
|---|---|---|---|
| CIFAR-10 | Coarse | × | × |
| CIFAR-10N-worse | Coarse | √ | × |
| CIFAR-10I | Coarse | × | √ |
| Oxford-IIIT Pet | Fine | × | √ |
| Oxford-IIIT Pet-N | Fine | √ | √ |

# **Single-Model Study**: Setting*

*See paper for details

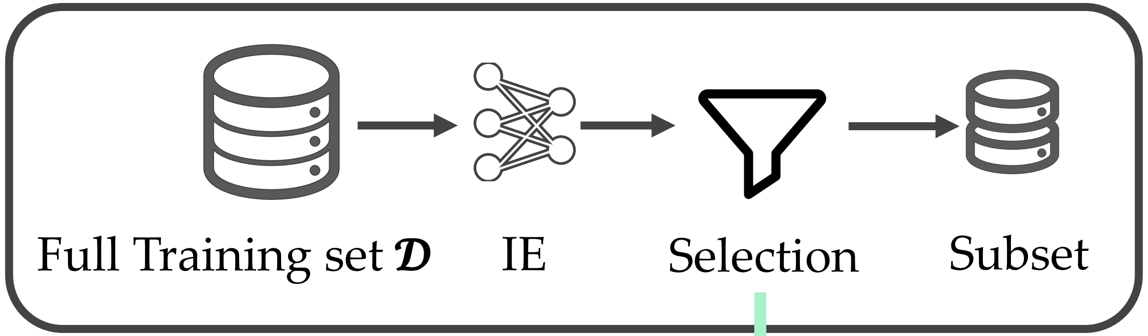
- **5** datasets × **3** types of IEs × **4** selection methods × **3** sampling rates



# **Single-Model Study**: Setting*

*See paper for details

- **5** datasets × **3** types of IEs × **4** selection methods × **3** sampling rates



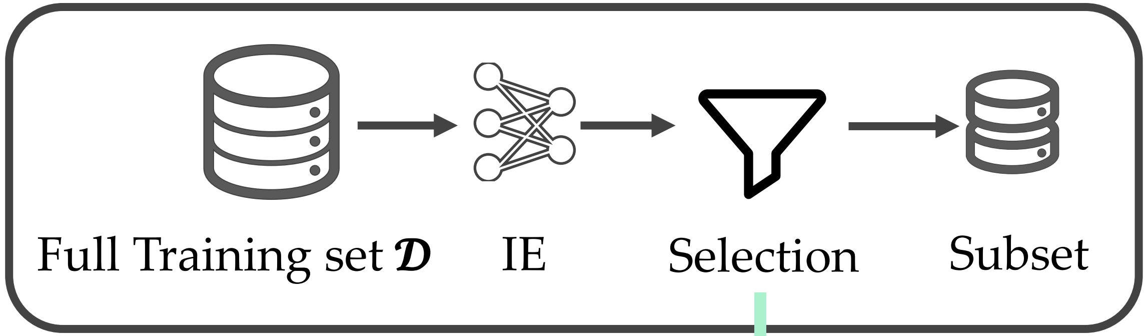

Feature-based subset selection methods
- Graph Cut (GC) [1]
- K-Center Greedy (KCG) [2]
- Moderate_DS (MDS) [3]
- MIN

[1] Iyer, R., and et al. Submodular combinatorial information measures with applications in machine learning. In Algorithmic Learning Theory. PMLR 2021.
[2] Sener, O., and et al. Active learning for convolutional neural networks: A core-set approach. ICLR 2018.
[3] Xia, X., et al. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. ICLR, 2023.

# **Single-Model Study**: Setting*

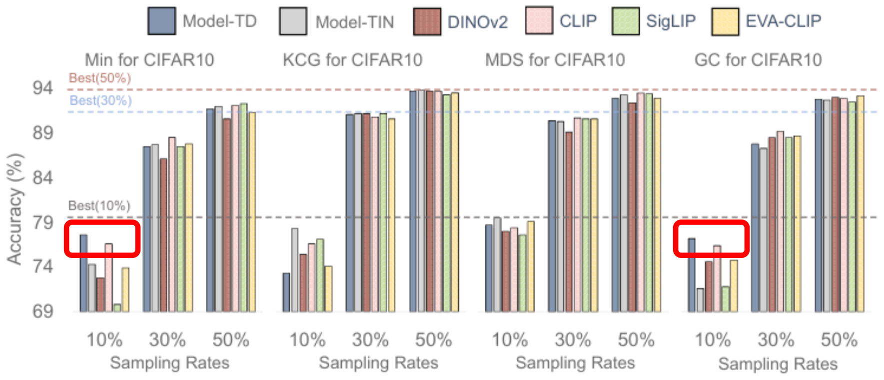
*See paper for details

- **5** datasets × **3** types of IEs × **4** selection methods × **3** sampling rates

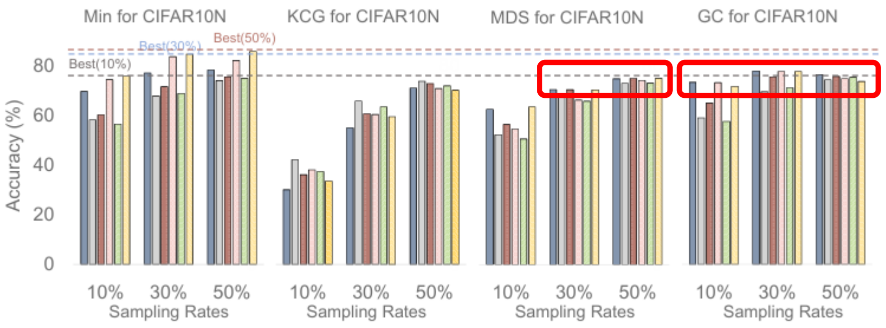



Sampling rates
- 10%
- 30%
- 50%

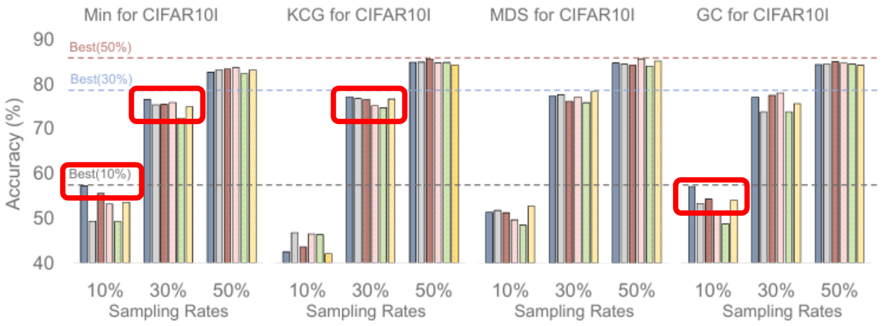
# **Single-Model Study:** FM ≠ Always Better



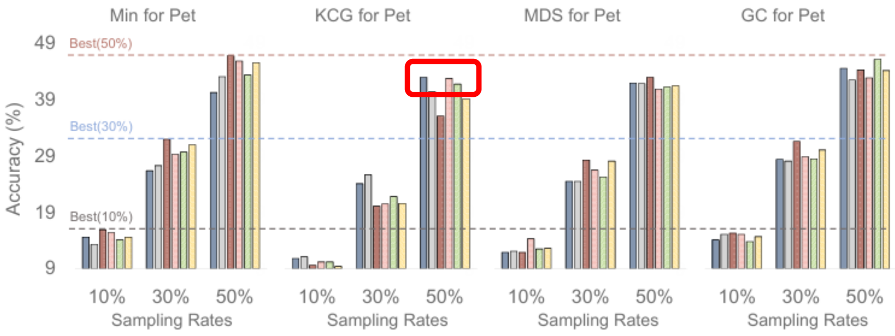

(a) Subset selection on CIFAR-10



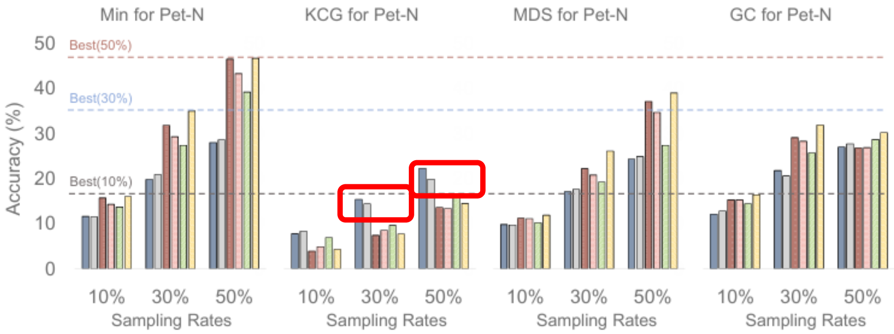

(b) Subset selection on CIFAR-10N-Worse (CIFAR-10N)




(c) Subset selection on CIFAR-10-imbalance (CIFAR-10I)




(d) Subset selection on Oxford-IIIT Pet (Pet)

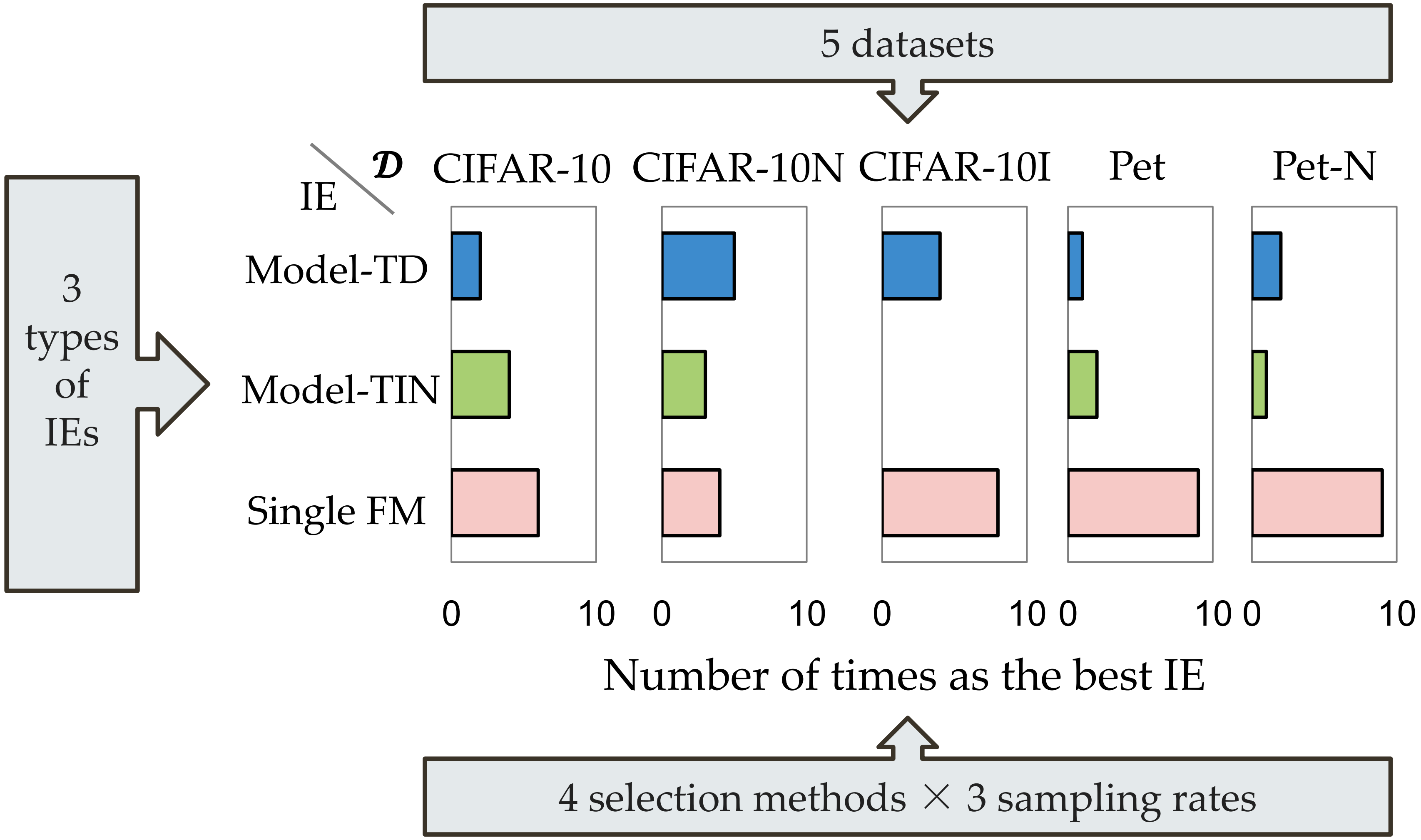



(e) Subset selection on Oxford-IIIT Pet with 20% symmetric label noise

- FMs do not always outperform traditional IEs.

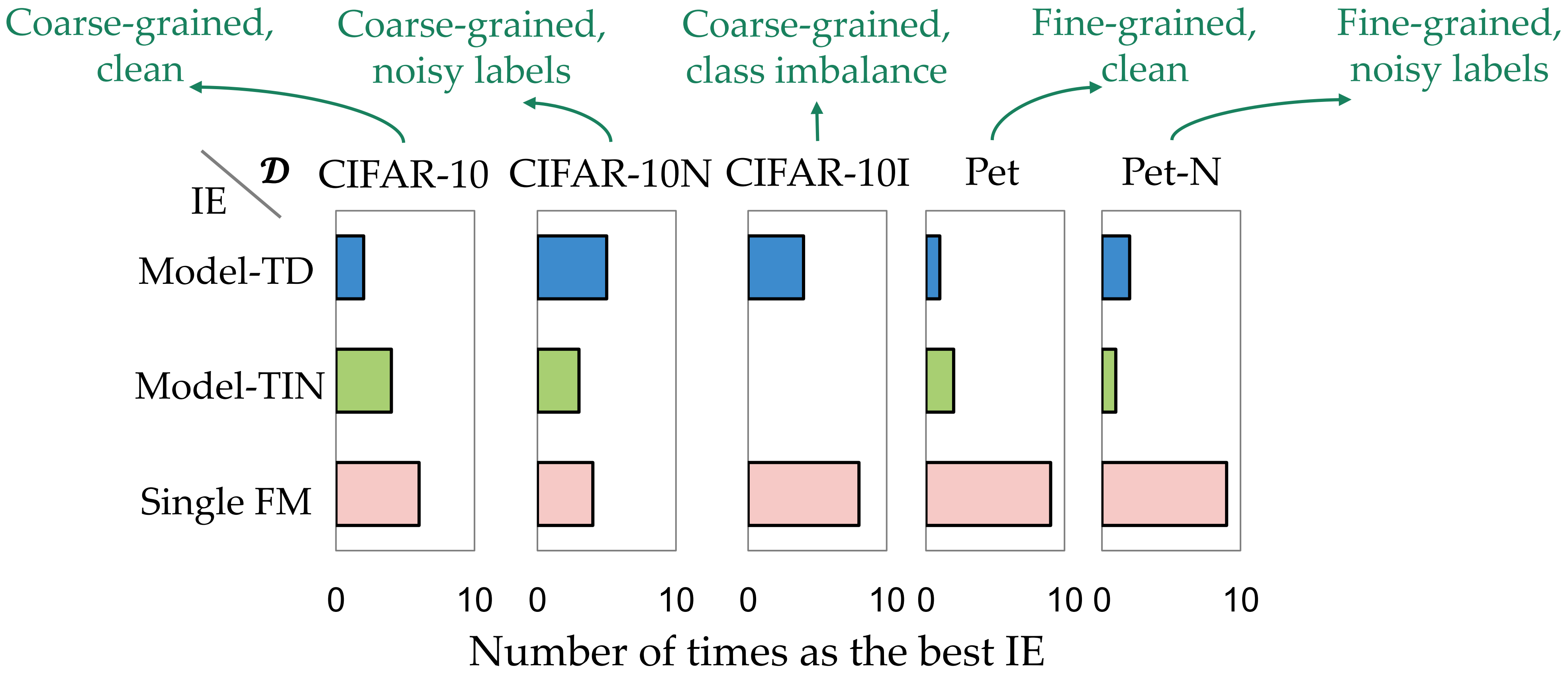
# Single-Model Study

**Best Extractor Frequency**, capturing how consistently an extractor is preferred



# Single-Model Study

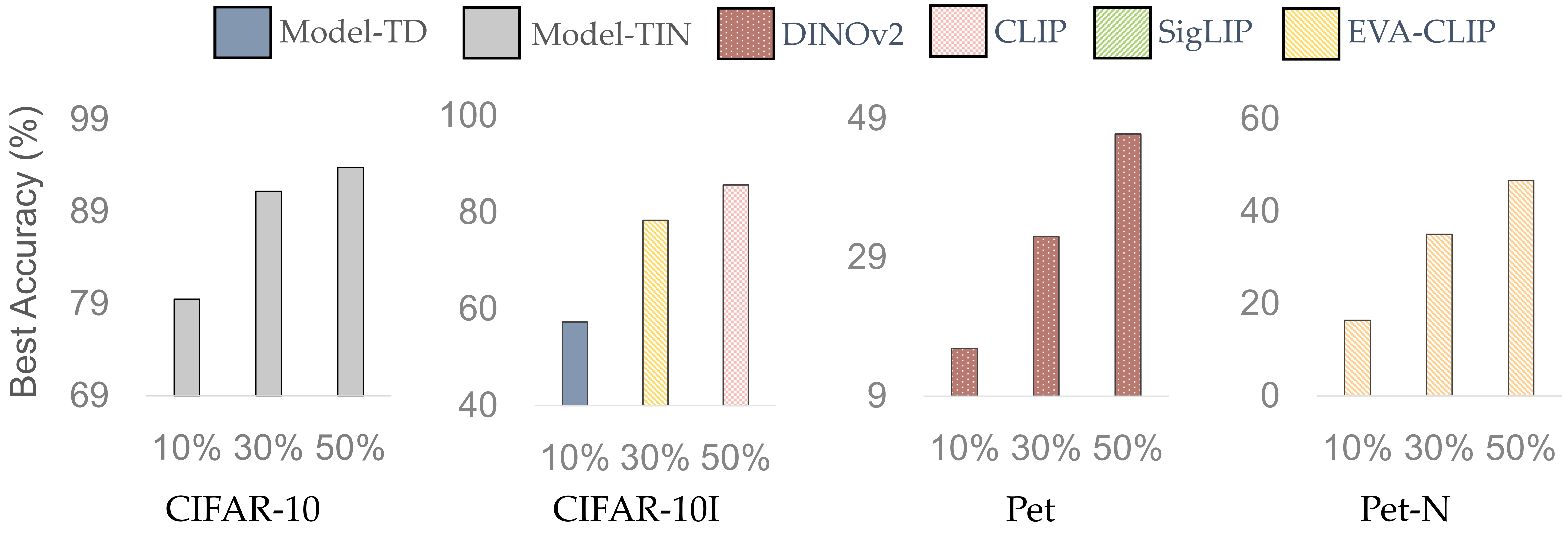
**Best Extractor Frequency**, capturing how consistently an extractor is preferred




- The single FM is preferred in only 4 out of 12 settings on CIFAR-10N, highlighting that **its advantage on noisy, coarse-grained data is limited and unstable**.
- On datasets like **CIFAR-10, CIFAR-10I, Pet, and Pet-N**, the **single FM is consistently preferred** over traditional IEs, with up to 9 out of 12 settings on the fine-grained datasets.

# Single-Model Study

**Performance Dominance**, examining which extractor achieves the best result at each sampling rate and their peak performance potential




- CIFAR-10: ✗ **No** Single FM wins (any rate)
- CIFAR-10I: ✓ Single FM wins at 30%, 50%; ✗ at 10%
- Pet / Pet-N: ✓ Single FM wins at all sampling rates

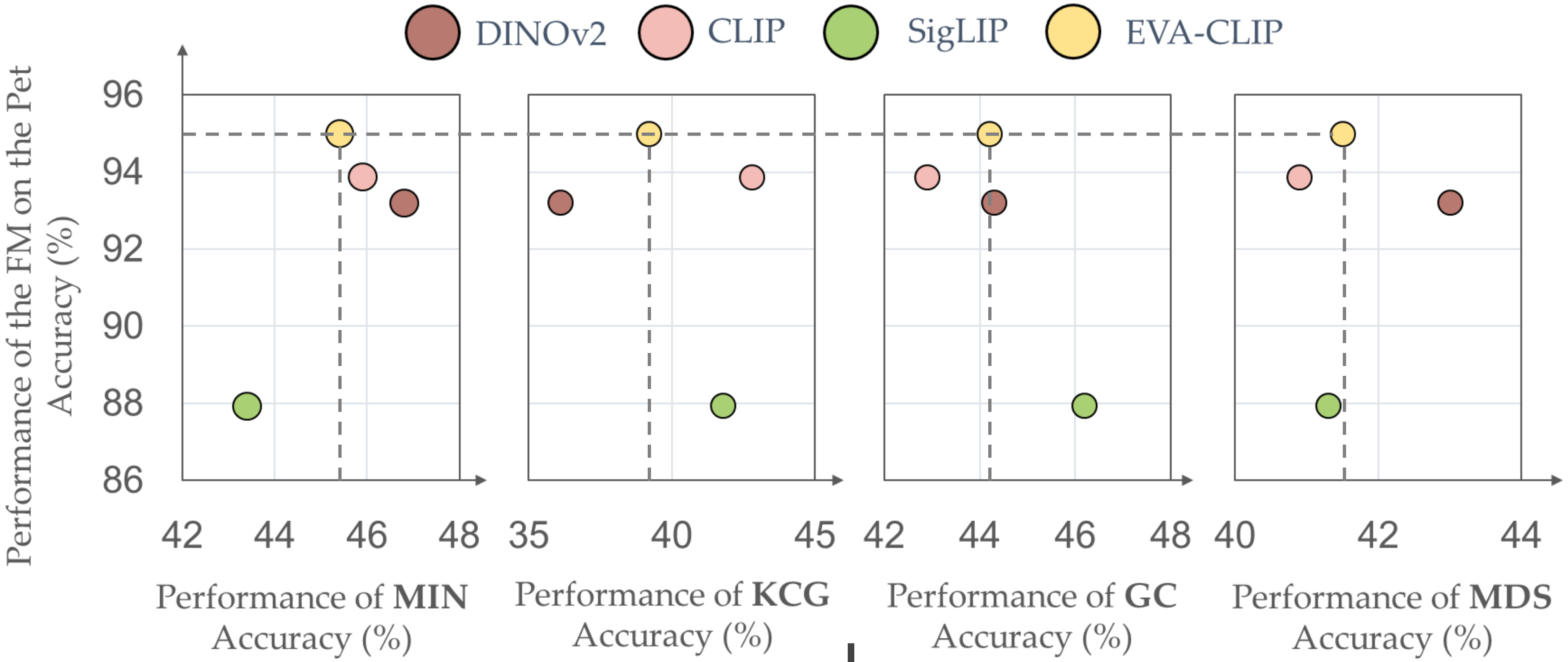
# **Single-Model Study**: When Do FMs Help Subset Selection?

- FMs significantly and consistently outperform traditional IEs for subset selection on fine-grained datasets (both clean and noisy).
- In contrast, FMs show limited or unstable advantages on coarse-grained datasets—especially when noisy labels are present, as in CIFAR-10N.

# **Single-Model Study**: Not All FMs Perform Equally Well As IE

**Observation 3**: Different FMs perform differently for subset selection, and the superior performance of FMs on downstream classification does not guarantee better subset selection effects.

Directly test the FM on the downstream task



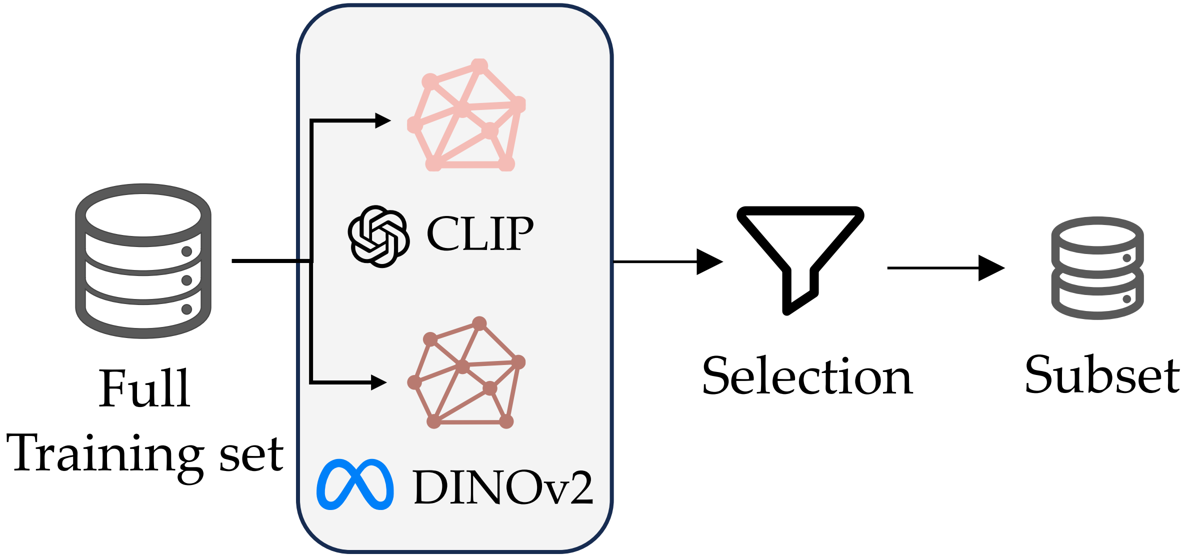

Effectiveness of each FM as an IE

# Single-Model Study:

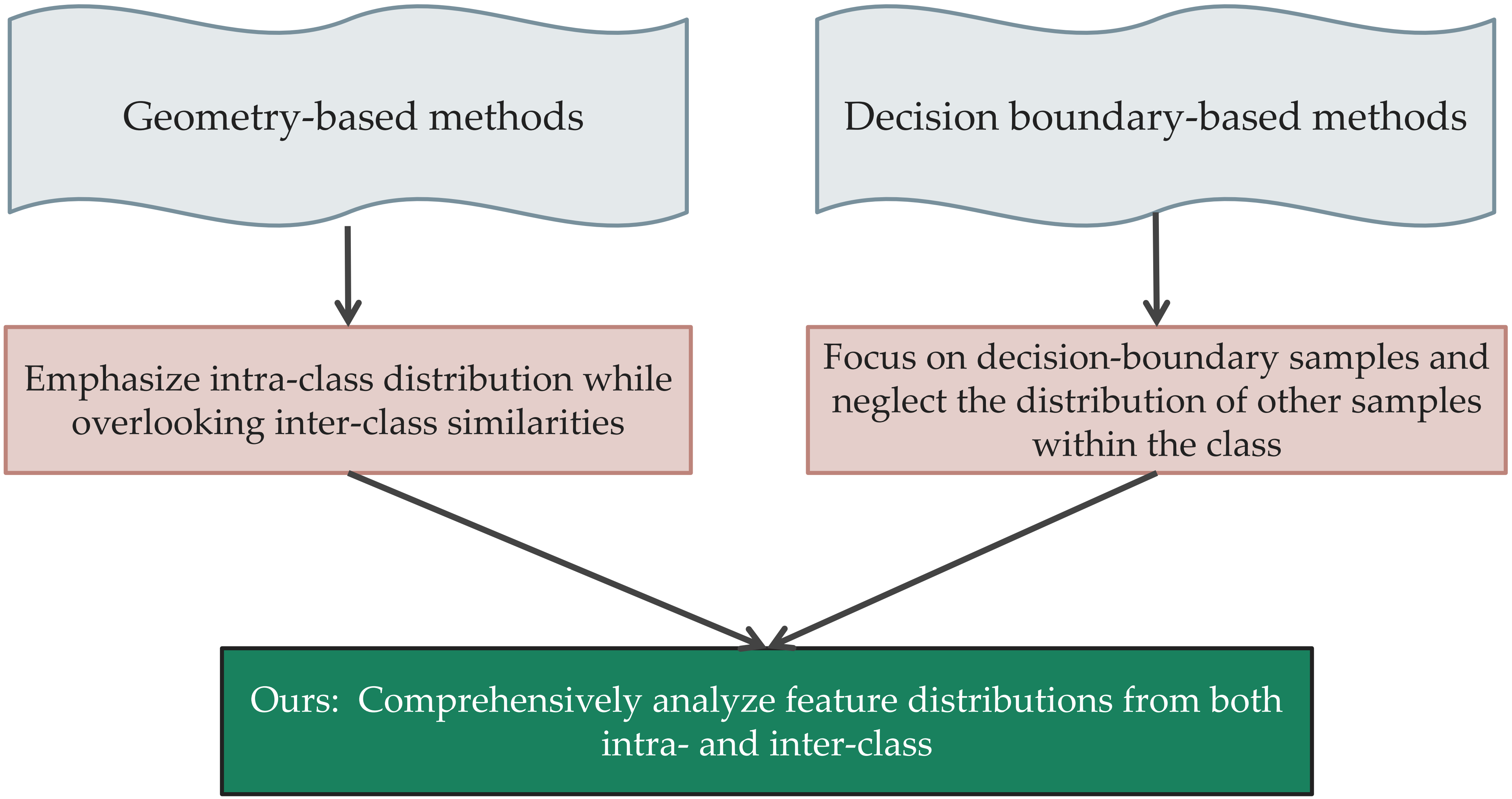
## Question

Can we combine the strengths of multiple FMs to explore the boundary of FM-based subset selection on fine-grained datasets?

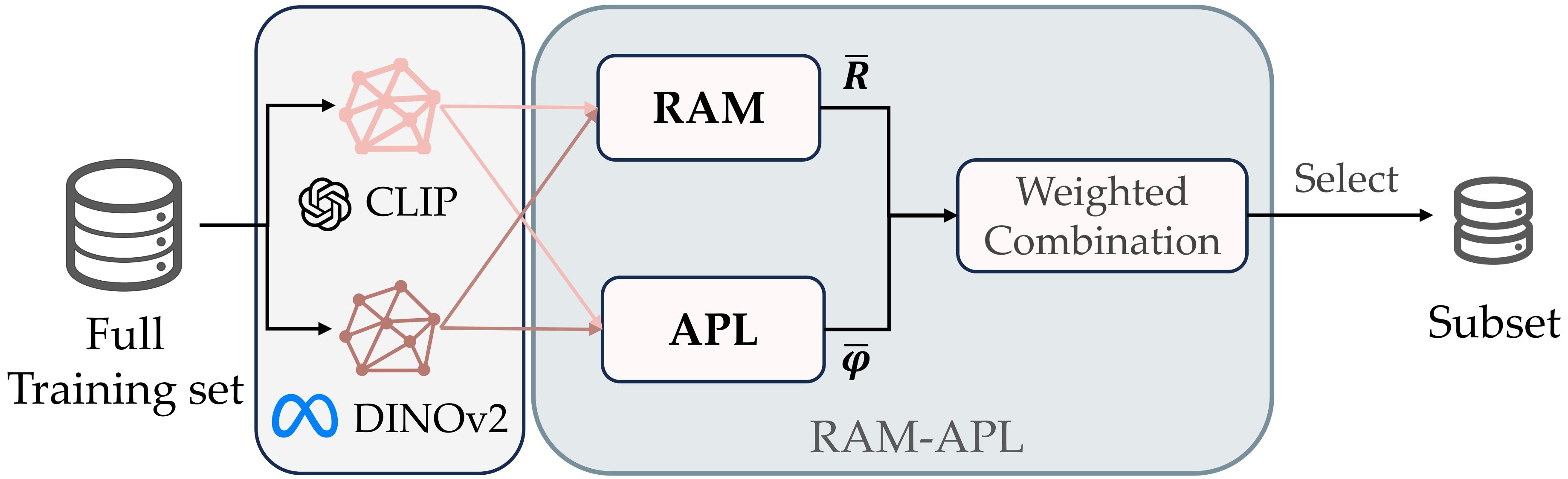
# **Proposed Method:** Multi-FM-based Subset Selection



# Conventional feature-based Subset Selection



# **Proposed Method:** RAM-APL

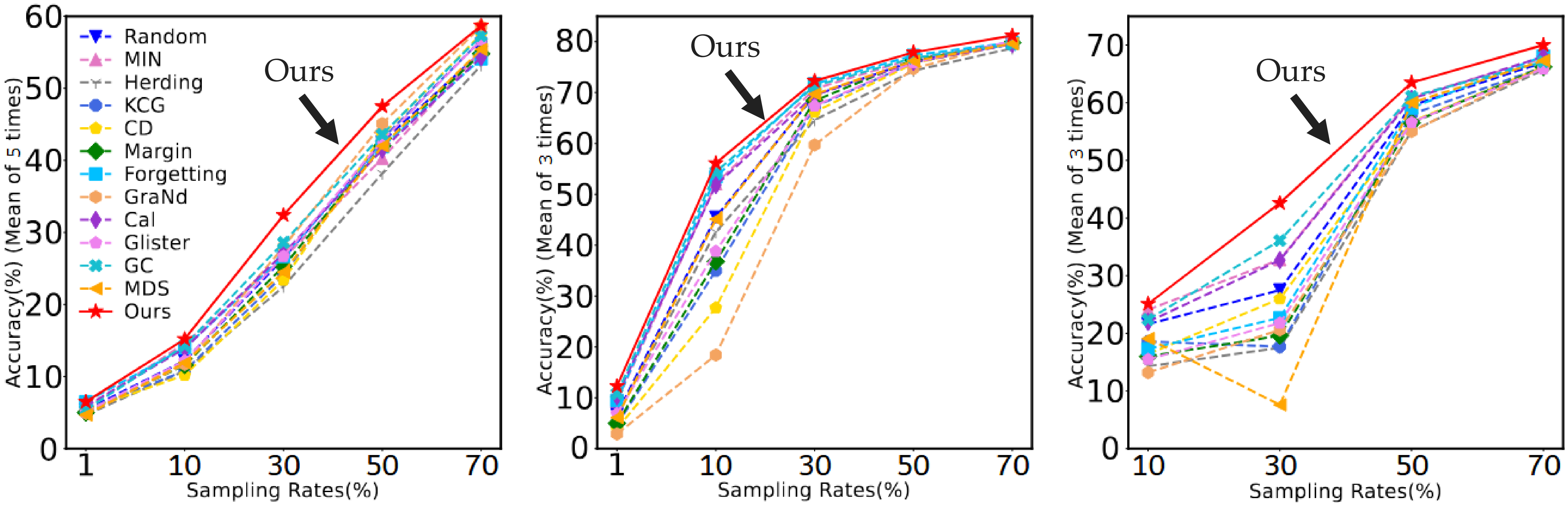



- **RAM (RAnking Mean):**
  - Aligns features from different FMs by mapping them into a unified distance ranking space;
  - Measures sample representativeness by averaging a sample's intra-class distance rank across multiple FMs.
- **APL (Accuracy of Pseudo Labels):**
  - Aligns features from different FMs by mapping them into a shared pseudo-label confidence space;
  - Averages pseudo-label accuracy across FMs to capture inter-class ambiguity.
- **RAM-APL:** A unified strategy that jointly evaluates representativeness (intra-class) and hardness (inter-class) by leveraging diverse FM perspectives.

# **Experimental Results:** Comparison with Baselines

- Our method outperformed all 12 subset selection baselines at each sampling rate.

Higher class complexity




(a) Oxford-IIIT Pet (b) Food-101 (c) CUB-200-2011

[!] All subset selection baselines follow the traditional pipeline.

# Experimental Results

Table. Comparison of the performance of our method using different numbers of foundation models as information extractors. Here, "D", "C", "S" and "E" represent DINOv2, CLIP, SigLIP, EVA-CLIP, respectively.

| | IE | | | | Sampling rates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | D | C | S | E | 1% | 10% | 30% | 50% | 70% | Overall Mean |
| Single | ● | ○ | ○ | ○ | 5.9±0.3 | 15.4±1.1 | 31.6±2.3 | 47.7±1.1 | 57.9±4.1 | 158.5 |
| | ○ | ● | ○ | ○ | 5.7±0.4 | 15.0±0.2 | 27.9±1.2 | 43.6±1.9 | 57.0±0.4 | 149.2 |
| | ○ | ○ | ● | ○ | 6.6±0.3 | 14.1±1.0 | 28.8±1.1 | 43.9±1.7 | 55.1±2.6 | 148.5 |
| | ○ | ○ | ○ | ● | 5.4±0.3 | 15.0±0.6 | 30.2±2.5 | 44.4±2.3 | 56.6±1.8 | 151.6 |
| Two | ● | ● | ○ | ○ | 6.5±0.4 | 15.2±1.2 | 32.4±2.9 | 47.5±1.9 | **58.7±2.2** | **160.3** |
| | ● | ○ | ● | ○ | 5.9±0.3 | 16.2±0.1 | 31.4±3.2 | 45.0±1.3 | 58.6±1.2 | 157.1 |
| | ● | ○ | ○ | ● | 6.0±0.6 | 16.0±0.9 | **35.8±2.9** | 46.5±1.8 | 54.9±3.5 | **159.3** |
| | ○ | ● | ● | ○ | 6.4±0.2 | 15.1±0.4 | 29.8±1.6 | 45.9±1.3 | 56.2±2.7 | 153.4 |
| | ○ | ● | ○ | ● | 5.9±0.3 | 15.5±0.7 | 31.4±1.7 | 44.2±2.2 | 55.9±1.8 | 152.9 |
| | ○ | ○ | ● | ● | **6.7±0.4** | 16.2±0.6 | 34.7±0.3 | 45.7±0.8 | 56.6±2.4 | **159.9** |
| Three | ● | ● | ● | ○ | 6.2±0.8 | 15.6±0.5 | 33.2±1.4 | **48.3±1.1** | 57.6±0.1 | **160.9** |
| | ● | ● | ○ | ● | 6.0±0.4 | **17.5±1.0** | 35.2±1.8 | 47.9±1.5 | 55.6±2.1 | **162.2** |
| | ● | ○ | ● | ● | 6.1±0.3 | 16.8±0.6 | 34.4±2.1 | 47.0±2.0 | 55.1±1.6 | **159.4** |
| | ○ | ● | ● | ● | 6.1±0.2 | 16.1±0.3 | 33.9±1.4 | 46.8±1.5 | 55.1±0.5 | 158.0 |
| Four | ● | ● | ● | ● | 6.5±0.2 | 16.8±1.1 | 34.0±2.7 | 46.3±0.5 | 56.9±1.1 | **160.5** |

- Combining multiple FMs can yield better overall performance than any single model.

# Experimental Results

Table. Comparison of the performance of our method using different numbers of foundation models as information extractors. Here, “D”, “C”, “S” and “E” represent DINOv2, CLIP, SigLIP, EVA-CLIP, respectively.

| IE | | | | Sampling rates | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| D | C | S | E | 1% | 10% | 30% | 50% | 70% | Overall Mean |
| ● | ○ | ○ | ○ | 5.9±0.3 | 15.4±1.1 | 31.6±2.3 | 47.7±1.1 | 57.9±4.1 | 158.5 |
| ○ | ● | ○ | ○ | 5.7±0.4 | 15.0±0.2 | 27.9±1.2 | 43.6±1.9 | 57.0±0.4 | 149.2 |
| ○ | ○ | ● | ○ | 6.6±0.3 | 14.1±1.0 | 28.8±1.1 | 43.9±1.7 | 55.1±2.6 | 148.5 |
| ○ | ○ | ○ | ● | 5.4±0.3 | 15.0±0.6 | 30.2±2.5 | 44.4±2.3 | 56.6±1.8 | 151.6 |
| ● | ● | ○ | ○ | 6.5±0.4 | 15.2±1.2 | 32.4±2.9 | 47.5±1.9 | **58.7±2.2** | 160.3 |
| ● | ○ | ● | ○ | 5.9±0.3 | 16.2±0.1 | 31.4±3.2 | 45.0±1.3 | 58.6±1.2 | 157.1 |
| ● | ○ | ○ | ● | 6.0±0.6 | 16.0±0.9 | **35.8±2.9** | 46.5±1.8 | 54.9±3.5 | 159.3 |
| ○ | ● | ● | ○ | 6.4±0.2 | 15.1±0.4 | 29.8±1.6 | 45.9±1.3 | 56.2±2.7 | 153.4 |
| ○ | ● | ○ | ● | 5.9±0.3 | 15.5±0.7 | 31.4±1.7 | 44.2±2.2 | 55.9±1.8 | 152.9 |
| ○ | ○ | ● | ● | **6.7±0.4** | 16.2±0.6 | 34.7±0.3 | 45.7±0.8 | 56.6±2.4 | 159.9 |
| ● | ● | ● | ○ | 6.2±0.8 | 15.6±0.5 | 33.2±1.4 | **48.3±1.1** | 57.6±0.1 | 160.9 |
| ● | ● | ○ | ● | 6.0±0.4 | **17.5±1.0** | 35.2±1.8 | 47.9±1.5 | 55.6±2.1 | 162.2 |
| ● | ○ | ● | ● | 6.1±0.3 | 16.8±0.6 | 34.4±2.1 | 47.0±2.0 | 55.1±1.6 | 159.4 |
| ○ | ● | ● | ● | 6.1±0.2 | 16.1±0.3 | 33.9±1.4 | 46.8±1.5 | 55.1±0.5 | 158.0 |
| ● | ● | ● | ● | 6.5±0.2 | 16.8±1.1 | 34.0±2.7 | 46.3±0.5 | 56.9±1.1 | 160.5 |

- DINOv2+CLIP achieves the best trade-off between efficiency and accuracy (**Our default setting**);

# Takeaways

- This work conducts, for the first time, a comprehensive analysis of the strengths and limitations of foundation models versus traditional information extractors (IEs) in subset selection. We find that
    1. Foundation models consistently outperform traditional IEs on fine grained datasets;
    2. This advantage diminishes particularly on coarse-grained datasets with noisy labels.

- The multi-FM-based subset selection method RAM-APL outperforms all baselines under different subset rates.

**Thank you so much for listening !**

**Visit our poster at East Exhibition Hall A-B #E-1912**

More details，please email wanzjwhu@whu.edu.cn

Paper Github