Mapping the Concept Landscape: Structural Perception of Global Distributions for Transparent Data Pruning

Dongyue Wu[1,3*] and Tao Ma[1,2*]

[1] State Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[3] Ant Group
{dongyue_wu, u202215201}@hust.edu.cn

Abstract. Existing data pruning methods predominantly rely on high-dimensional feature embeddings to measure sample importance. However, these compressed vectors often obscure fine-grained semantic interactions, leading to suboptimal coverage of rare semantic concepts in the pruned subsets. In this paper, we propose Mapping the Concept Landscape (MCL), a novel structural perception framework for transparent data pruning. Instead of abstract embeddings, we represent each image-caption pair as an explicit sample-level graph comprising entities, events, and attributes. By integrating these individual graphs into a comprehensive dataset-level graph, we characterize the global distribution of semantic concepts and quantify their rarity across the entire corpus. Based on this structured perception, we develop a greedy concept-coverage maximization algorithm that iteratively selects samples to maximize the marginal gain of high-value, under-represented concepts. Experimental results on various benchmarks demonstrate that our method not only achieves superior pruning efficiency compared to state-of-the-art methods but also provides a transparent and interpretable audit trail for the selection process.

Keywords: Data pruning · Training efficiency · Deep learning

1 Introduction

The rapid expansion of web-scale vision–language datasets, such as LAION [42] and CC12M [6], has been a key driver behind the recent success of large vision-language models [9, 38]. However, this growth also introduces practical challenges. Training on massive datasets incurs prohibitive storage and computational costs, while the automated data collection process inevitably introduces a large amount of redundant, low-quality, and noisy samples. As a result, data

* Equal contribution.

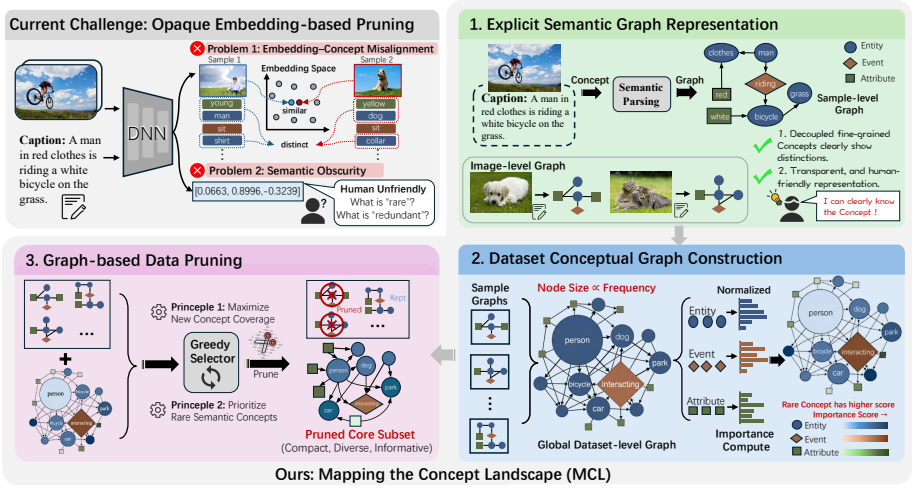


Fig. 1: Illustration of the challenge of data pruning methods and the overview of our proposed Mapping the Concept Landscape (MCL) framework for transparent data pruning. MCL aims to maximize semantic coverage based on concept graphs. The node size is positively related to the number of samples a concept appears in the dataset.

pruning [35, 56], has become a crucial technique for improving data efficiency, aiming to select a compact yet representative subset that preserves model performance under strict data budgets.

At its core, dataset pruning is an optimization problem: given a limited budget, which subset of samples should be retained to best preserve the information of the original dataset? Most existing pruning methods address this problem at the level of sample-wise representations, selecting data based on loss value [33, 35, 37], uncertainty [3, 8, 15, 39], learning difficulty [46, 49], or influence [20] measured in a compressed sample-level embedding space. Other works [16,22,44,51,58] attempt to preserve dataset diversity by selecting samples according to similarity or distance between embeddings. These methods implicitly assume that geometric similarity in the embedding space faithfully reflects semantic redundancy in the underlying data.

However, this assumption leads to a fundamental *Embedding–Concept Misalignment*. While embedding similarity measures proximity between samples in a learned geometric space, semantic diversity is inherently compositional and structured at the concept level. Consequently, redundancy in the embedding space does not reliably correspond to redundancy in semantic concepts. Samples that appear redundant in representation space may still contain distinct fine-grained concepts whereas geometrically diverse samples may largely overlap in their conceptual content. This mismatch arises because rich multi-concept semantics are collapsed into a single coupled vector representation. Hence, embedding-based pruning methods primarily optimize geometric diversity rather than true semantic coverage, and may systematically discard low-

fraction yet semantically important concepts even when the retained subset appears well-distributed in embedding space.

Moreover, compressing each image–text pair into a dense numerical embedding introduces another practical limitation: a lack of interpretability at the dataset level. Pruning decisions are made in an opaque embedding space, making it unclear for human which semantic concepts are preserved or discarded. While such representations may be effective for model optimization, they are difficult for humans to interpret. As a result, practitioners have limited ability to audit, inspect, and refine pruning outcomes or reason about the resulting semantic distribution of the whole dataset.

Together, these limitations stem from treating samples as indivisible units and obscuring their internal semantic structure, motivating a shift from sample-level pruning to concept-level coverage. In this paper, we propose *Mapping the Concept Landscape* (MCL), a graph-structured framework for multimodal dataset pruning that explicitly models and optimizes concept-level coverage. Instead of representing each sample as an opaque vector, MCL maps each image–text pair to an interpretable sample-level semantic graph composed of atomic concepts, including Entities, Events, and Attributes. While concept extraction may be imperfect at the individual sample level, MCL operates on aggregated concept statistics across the dataset, where relative frequencies and coverage patterns dominate pruning decisions. By aggregating sample-level graphs, MCL constructs a dataset-level concept graph that provides a transparent, human-friendly view of the dataset's semantic landscape, enabling direct inspection of concept distributions, redundancies, and gaps. Based on this dataset-level representation, pruning is formulated as a coverage maximization problem over semantic concepts. Concepts that appear frequently contribute diminishing additional value, whereas underrepresented concepts provide higher marginal benefit when introduced into the selected subset. This naturally leads to a greedy selection strategy that prioritizes samples according to their marginal contribution, namely the extent to which they introduce previously uncovered, high-value concepts. Importantly, emphasizing rare concepts in this manner does not imply over-sampling the long tail. Since concept coverage is a monotone and saturating objective, the marginal gain of repeatedly selecting samples containing the same concepts quickly diminishes, enabling the greedy procedure to balance redundancy reduction with the preservation of dominant semantics. In summary, our main contributions are three-fold:

- We propose MCL, a data pruning framework that moves beyond opaque embedding-based criteria by representing data using structured concept graphs, enabling effective pruning and human-interpretable dataset analysis.
- We introduce a greedy pruning algorithm that selects samples based on their marginal contribution to concept coverage, prioritizing underrepresented concepts and reducing redundant and dominant semantics.
- Extensive experiments demonstrate that MCL achieves state-of-the-art pruning efficiency: using less than 10% of training time and selected data, our method attains performance comparable to full-data baselines.

2 Related Work

Dataset Pruning and Coreset Selection. Dataset pruning and coreset selection aim to identify a compact subset of training data that preserves the performance of the full dataset. Early studies explored this idea in contexts such as active learning [4, 43], noisy learning [34], and continual learning [2, 40]. Many approaches estimate sample importance using signals derived from training dynamics [35, 49, 52], loss values [19, 30, 33, 35, 37], gradient [16, 35, 47, 56]. Other works focus on clustering or the distance to centers [46, 60] are used to select representative instances. These methods have demonstrated effectiveness in reducing dataset size while maintaining model performance. However, most existing techniques rely on compressed feature embeddings to define pruning metrics, treating each sample as an indivisible unit summarized by a single vector representation.

Data Selection for Multimodal Data. With the emergence of large vision–language models, data selection has recently attracted attention in visual instruction tuning. Several methods [7, 50, 55, 57, 58, 61] attempt to evaluate the usefulness of multimodal instruction data through model-driven signals. For example, Self-Filter [53] assigns scores to instruction data using a learned scoring network, while TIVE [29] estimates sample importance through gradient-based analysis of the target model. Other approaches leverage auxiliary models [57] or clustering techniques [22, 58] to identify important samples. Methods like ICONS [54] explores the size allocation across different instruction tasks.

Diversity and Coverage-Based Data Selection. A line of research emphasizes maintaining diversity or coverage in the selected subset [16, 44, 48, 51, 58]. These approaches typically perform clustering or similarity-based sampling to ensure broader coverage of the data distribution. However, diversity is usually measured in embedding space, where complex samples are represented by single compressed vectors. Such representations often obscure the internal semantic structure of image–text pairs and make it difficult to reason about fine-grained concept coverage. In contrast, our method models datasets through explicit semantic concepts, enabling interpretable and fine-grained coverage-aware pruning.

3 Method

We present Mapping the Concept Landscape (MCL), a concept-centric framework for multimodal dataset pruning. Given a dataset of image–text pairs, MCL explicitly models the dataset's semantic structure at the concept level and selects a representative subset by maximizing fine-grained semantic coverage under a predefined budget. As illustrated in Fig. 1, MCL consists of the following key steps: First, each sample is mapped to a sample-level concept graph that captures its fine-grained semantic content using explicit and interpretable atomic concepts. Second, these sample-level graphs are aggregated to construct a dataset-level concept graph, which summarizes the global semantic distribution of the dataset. Third, based on concept frequencies, we define a concept

importance measure and formulate dataset pruning as a concept coverage maximization problem. Finally, we adopt a greedy selection strategy that iteratively selects samples according to their marginal contribution to the overall concept coverage. This formulation enables MCL to reduce redundancy while preserving diverse and semantically informative concepts, and simultaneously provides a transparent interface for understanding and auditing the pruned dataset.

3.1 Sample-level Concept Graph Construction

A central design choice of MCL is to represent each sample using an explicit, transparent, and fine-grained semantic representation rather than a single coupled and obscured embedding. To this end, we construct an image-level concept graph $G_i = (V_i, E_i)$ for each image–text pair $x_i = (I_i, T_i)$. The caption T_i can be either the original accompanying caption or a generated caption produced by an off-the-shelf vision–language model. Each node v in the nodes set V_i serves as the atomic semantic unit for subsequent dataset-level distribution modeling. A single node v represents an atomic concept, and edges in set E_i encode the concept co-occurrence and dependency relationships within the caption T_i. We also categorize each node to one of the three concept types:

- **Entities**, corresponding to nouns that denote objects or visual subjects;
- **Events**, corresponding to verbs that describe actions or interactions;
- **Attributes**, corresponding to adjectives or adverbs that specify properties, states, or modifiers of entities and events.

These three categories jointly capture the majority of fine-grained semantics commonly present in image–text descriptions, including objects, actions, attributes, and their contextual relations. Please refer to Appendix for visualizations. Compared to representing a sample as a single embedding vector, this graph-based representation explicitly decomposes the sample into interpretable semantic components, enabling fine-grained reasoning at the concept level.

The specific procedure used to extract concepts and construct the image-level graph is not a contribution of this work and can be implemented using standard linguistic analysis tools. For example, we can either employ scene graph generation methods [21] based on I_i or simply extract the conceptual graph from T_i using dependency parsing methods [11, 12, 36]. We find that employing dependency parsing on these captions is not only computationally efficient but also effective at mapping a wide range of conceptual details. Therefore, we utilize the spaCy [17], a widely-adopted industrial-strength NLP library, to perform text-based dependency parsing on each x_i based on the text caption T_i. The process identifies and filters out stop words, retaining only semantically meaningful tokens. Based on the dependency tree and Part-of-Speech (POS) tags produced by spaCy, we construct the sample-level concept graph G_i. The image-level graph serves as a structured container that exposes the sample's internal semantic composition, which can later be aggregated and analyzed at the dataset level.

3.2 Dataset-level Concept Graph Construction

While image-level graphs capture fine-grained semantics for individual samples, our framework requires modeling the semantic coverage and concept distribution at the scale of the entire dataset. To enable such global analysis, we aggregate all image-level graphs into a unified dataset-level concept graph.

Formally, given a dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, we construct a dataset-level graph $G_{\mathcal{D}} = (V_{\mathcal{D}}, E_{\mathcal{D}})$ by taking the union of all concept nodes $V_D = \bigcup_{i=1}^{N} V_i$. Each node $v_j \in V_D$ is associated with a weight that counts the number of samples where the corresponding concept appears:

$$w(v) = \sum_{i=1}^{N} \mathbf{1}[v \in V_i]. \tag{1}$$

This weight reflects the empirical fraction of the concept in the dataset and serves as a proxy for its redundancy or rarity. The resulting dataset-level concept graph $G_{\mathcal{D}}$ provides an explicit summary of the dataset's semantic landscape. Concepts with large weights correspond to frequently occurring semantics, while low-weight nodes indicate rare and scarce concepts. Unlike embedding-based summaries, this representation of the whole dataset is directly interpretable by humans and enables straightforward inspection of concept distributions and semantic relations within the dataset. The constructed $G_{\mathcal{D}}$ is a summary structure which employs the graph to model the concept distribution, rather than an object to be pruned or conduct data selection. Please note, the actual objects to be pruned are still the individual samples x_i.

Crucially, MCL operates on aggregated concept statistics rather than individual sample annotations. Although concept extraction at the image level may be noisy or imperfect, such errors are amortized when aggregated across the dataset. Since pruning decisions are driven by relative concept frequencies and coverage patterns, the dataset-level concept graph provides a stable and robust foundation for defining concept importance and guiding data selection, which we describe in the following sections.

3.3 Concept Importance and Coverage Objective

Based on the dataset-level concept graph, we now formalize dataset pruning as a concept coverage maximization problem. We first define a concrete instantiation of concept node importance, and then introduce the resulting coverage objective.

Node Importance with Structural Diversity Awareness. We first define a concept importance function $\phi(v)$ for each concept node $v \in V_{\mathcal{D}}$. Intuitively, concepts that occur frequently tend to encode common and thus redundant semantics, whereas rare concepts are more likely to capture distinctive or underrepresented information. While $w(v)$ captures how common a concept is, it does not account for how the concept participates in the overall semantic structure of the dataset. A concept that appears rarely but only in isolated contexts

may be less informative than a rare concept that interacts with many other concepts through diverse relationships. To capture both aspects, we define concept importance by jointly considering semantic rarity and structural participation. Specifically, for each concept node v, we define its unnormalized node importance score $\tilde{\phi}(v)$ as follows

$$\tilde{\phi}(v) = \log\left(\frac{N_{type(v)}}{w(v)}\right) \cdot [1 + \log(d(v))], \tag{2}$$

where $N_{type(v)}$ denotes the total number of concept nodes within the same semantic category (Entity, Event, Attribute) as v, and $d(v)$ is the degree of v in $G_{\mathcal{D}}$. The first term is an inverse-fraction factor that assigns higher importance to insufficiently represented concepts, while the second term serves as a lightweight relational correction, promoting concepts that participate in richer semantic interactions. This design allows relational information to be incorporated without explicitly modeling edge importance, resulting in a simpler and more efficient formulation.

Category-wise Normalization. Concept nodes in the dataset graph belong to three semantic categories: Entity, Event, and Attribute. These categories naturally differ in scale and fraction distribution. To prevent any single category from dominating the importance scores, we perform normalization independently within each category. Formally, the final concept importance is defined as

$$\phi(v) = \frac{\tilde{\phi}(v) - \min_{u \in V_{type(v)}} \tilde{\phi}(u)}{\max_{u \in V_{type(v)}} \tilde{\phi}(u) - \min_{u \in V_{type(v)}} \tilde{\phi}(u) + \epsilon'}, \tag{3}$$

where $V_{type(v)}$ denotes the set of nodes belonging to the same category as v, and ϵ is a small constant for numerical stability. This normalization ensures that importance scores are comparable across different concept types, allowing rare events or attributes to contribute meaningfully alongside entity concepts.

Concept Coverage Function. Given a subset of selected samples $\mathcal{S} \subseteq \mathcal{D}$, we define its covered concept node set as

$$V_{\mathcal{S}} = \bigcup_{x_i \in \mathcal{S}}^{N} V_i, \tag{4}$$

where V_i denotes the set of concept nodes associated with sample x_i. The overall concept coverage function of $\mathcal{S}$ is then defined as

$$F(S) = \sum_{v \in V_{\mathcal{S}}} \phi(v). \tag{5}$$

This formulation explicitly accounts for semantic redundancy: once a concept is covered by the selected subset, additional samples containing the same concept do not further increase the objective value. Therefore, samples that introduce uncovered and structurally informative concepts yield higher marginal gains.

Dataset Pruning Objective. With the above definitions, dataset pruning can be formulated as the following constrained optimization problem:

$$\max_{S \subseteq D} F(S) \quad \text{s.t.} \quad |\mathcal{S}| \leq b, \tag{6}$$

where $b = (1-p) \cdot |\mathcal{D}|$ denotes the target number of the retained subset after data selection and p denotes the pruning ratio. By optimizing concept coverage under this objective, MCL prioritizes samples that jointly preserve rare, diverse, and relationally informative semantics, while maintaining a transparent and interpretable pruning criterion grounded in the dataset's concept structure.

3.4 Semantic Coverage Greedy Selection

With the proposed optimization objective Eq. 6, dataset pruning can be viewed as selecting a subset of samples whose associated concept sets collectively maximize semantic coverage under a fixed budget. This formulation naturally leads to an iterative greedy optimization strategy based on marginal coverage gain. At each step of the pruning process, the key question is which sample should be added to the current retained set in order to best improve the overall coverage. Since the objective function rewards the inclusion of previously uncovered concepts, the contribution of a candidate sample depends on the concepts that it introduces beyond those already covered.

Let $\mathcal{S}_t \subseteq \mathcal{D}$ denote the selected subset of samples of the t-th step. For any candidate samples $x \in \mathcal{D} \backslash \mathcal{S}_t$ with the corresponding concept graph $G_i = (V_i, E_i)$, its marginal gain with respect to the coverage objective is defined as

$$\Delta(x_i \mid \mathcal{S}) = F(\mathcal{S} \cup \{x_i\}) - F(\mathcal{S}) = \sum_{v \in V_i \backslash V_{\mathcal{S}}} \phi(v). \tag{7}$$

This marginal gain measures the boundary expansion of the current retained subset in the concept space: samples that introduce new, important, and hard-to-cover concepts receive higher scores, while samples whose concepts are already well covered contribute little or no gain. The greedy strategy then selects the sample with the largest marginal gain at each iteration:

$$x^* = \arg\max_{x \in \mathcal{D} \backslash \mathcal{S}_t} \Delta(x \mid \mathcal{S}_t), \tag{8}$$

which will be added to the selected subset $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{x^*\}$ for the next iteration. This procedure is repeated until the target budget is reached. Please refer to Algorithm 1 for the whole process.

The necessity of a greedy selection strategy stems directly from the nature of our coverage-based objective. Traditional pruning methods typically assign a static, independent importance score to each sample, which does not consider the selected subset as a whole. This scoring manner inevitably leads to select clusters of redundant samples that share the same high-value concepts. In contrast, our objective F inherently evaluate the overall concept coverage of the selected

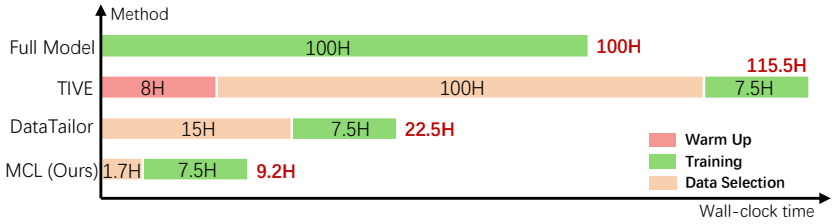


Fig. 2: Wall-clock time comparison with recent data pruning methods [29,58]. Results are collected on a device with four NVIDIA RTX 3090 GPUs.

subset. However, Eq. 6 is typically an NP-hard optimization problem, and it is unacceptable to compute the set-dependent $F(\mathcal{S})$ for $\binom{b}{|\mathcal{D}|}$ possible choice. To solve Eq. 6 more efficiently, we resort to such an iterative greedy algorithm. Our greedy strategy evaluates the dynamic marginal gain of each candidate, specifically its ability to introduce previously uncovered concepts. Because the value of a sample is explicitly conditioned on the state of the currently retained subset, the greedy rule continuously targets the data semantic frontier. This dynamic evaluation ensures that every newly selected sample maximally complements, rather than repeats, the existing concepts, which is the key to eliminating redundancy in large, repetitive datasets. The greedy selection rule depends only on set differences and concept importance scores, allowing efficient incremental updates. In practice, we further accelerate the pruning process by partitioning the dataset into chunks and evaluating marginal gains in parallel across candidate samples, making the approach efficient and scalable to web-scale datasets. For examples, on LLaVA-1.5-mix-665k [27] dataset, we partition the whole dataset into eight chunks and conduct greedy data pruning in parallel. As reported in the Fig. 2, the proposed MCL only introduce a 1.7-hour overhead while the SOTA methods like TIVE [29] spend days to conduct data pruning. The reported overheads include both the graph building and the selection time.

4 Experiment

4.1 Datasets & Backbone

To comprehensively evaluate the effectiveness and generalization of our dataset pruning framework, we conduct experiments on both multimodal vision–language instruction tuning (VIT) and object detection, a classic pure vision task. This dual setting allows us to assess whether our method can generalize to fundamentally different learning paradigms for multimodal and only image data.

Vision–Language Instruction Tuning. For multimodal instruction tuning, we adopt the LLaVA-1.5-mix-665k dataset [28]. This dataset contains approximately 665K instruction-following samples collected from 12 vision–language tasks, covering a wide range of visual reasoning and perception abilities. Its scale, diversity, and heavy semantic redundancy make it a representative benchmark

Algorithm 1 Greedy Concept Coverage Pruning (MCL)

Input: Dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, concept graph set $\{G_i = (V_i, E_i)\}$ for each $x_i \in \mathcal{D}$, dataset-level concept graph $G_{\mathcal{D}}$, concept node importance $\phi(v)$ and target budget $b = (1-p)N$.
Output: Retained sample subset $\mathcal{S}$.

$\mathcal{S}_0 \leftarrow \emptyset$, $V_{\mathcal{S}_0} \leftarrow \emptyset$. ▷ Initialize
for t **in** $[0, b]$ **do**
  for each $x_i \in \mathcal{D} \setminus \mathcal{S}_t$ **do**
    Get $\phi(v)$ based on $G_{\mathcal{D}}$ for all nodes in V_i.
    Get the marginal gain $\Delta(x_i \mid \mathcal{S}_t) \leftarrow \sum_{v \in V_i \setminus V_{\mathcal{S}_t}} \phi(v)$.
  end for
  Pick the optimal sample $x^* \leftarrow \text{argmax}_x \, \Delta(x \mid \mathcal{S}_t)$.
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{x^*\}$. ▷ Update retained subset
  $V_{\mathcal{S}_{t+1}} \leftarrow V_{\mathcal{S}_t} \cup V_{x^*}$.
end for
return $\mathcal{S} \leftarrow \mathcal{S}_b$

for evaluating dataset pruning strategies in large-scale multimodal training. We employ the LLaVA-v1.5-7B [28] model as our backbone.

Object Detection. To further examine the robustness and task-agnostic nature of our approach, we additionally evaluate MCL on the COCO 2017 object detection dataset [26]. COCO 2017 contains 118k training images and 5k validation images annotated with 80 object categories. There are 7 instances per image on average, ranging from small to large on the same images. We conduct all experiments using DETR [5] with a ResNet-50 [14] backbone. Compared to single-label image classification dataset such as ImageNet [10], object detection requires both category recognition and precise spatial localization, and thus relies on richer and more fine-grained visual semantics. This makes COCO more suitable for assessing whether preserving semantic coverage can effectively support complex visual task. Although object detection is a purely vision-based task, COCO 2017 provides an additional advantage: each training image is paired with five human-written captions. These captions offer a compact, high-level semantic abstraction of visual content, enabling efficient concept extraction.

Together with the VIT setting, experiments on COCO 2017 demonstrate that our coverage-based pruning strategy generalizes beyond multimodal instruction learning and remains effective in a classical vision task with substantially different supervision and learning paradigm.

Evaluation Benchmarks. For VIT, we evaluate the pruned methods on a set of representative benchmarks covering multimodal perception, reasoning, and general visual question answering. Specifically, we include MME [13] for perceptual and cognitive evaluation, SEED-Bench-Image [23] for comprehensive multimodal assessment across multiple dimensions, POPE [25] for hallucination analysis, and MMMU [59] for challenging scientific reasoning. To assess general

Table 1: Comparison to state-of-the-art methods on LLaVA-v1.5-mix-665k dataset with 50k (7.5%) and 133k (20%) selected data. Our results are shown in the gray block. The best and the second best results are in **bold** and underlined, respectively. We all use the LoRA model due to limited resources. † denotes our implementation.

Benchmarks	Valid Data	MME-P	MME-C	SEED-I	POPE	MMMU	SQA	GQA	TextQA	Rel.
Baseline	665k	1476.9	267.9	67.4	86.4	32.8	70.0	63.0	58.2	100.0%
Random	50k	1387.5	287.5	59.7	**85.7**	32.2	68.4	55.0	53.1	95.7%
Length	50k	1357.0	265.7	47.0	82.6	33.9	60.9	55.5	45.2	89.1%
EL2N [35]	50k	1077.3	252.5	59.3	80.8	33.6	71.0	**61.0**	41.7	90.1%
GradN [35]	50k	1275.4	303.6	58.3	75.7	32.4	70.9	44.9	46.0	90.5%
IFD [24]	50k	1113.4	301.8	55.1	76.7	33.0	48.2	41.9	43.6	83.6%
InsTag [31]	50k	1317.1	**345.0**	57.4	82.1	**34.0**	69.3	52.5	53.3	97.0%
LESS [55]	50k	1344.8	281.8	61.2	79.4	33.0	71.0	53.4	52.0	94.4%
Self-Filter† [53]	50k	1331.2	262.0	58.2	78.1	31.4	68.9	52.8	50.1	91.0%
TIVE† [29]	50k	1434.8	291.5	**61.6**	84.9	33.3	**71.2**	56.3	52.0	97.2%
DataTailor† [58]	50k	1447.4	322.3	60.6	82.4	33.9	70.4	57.1	52.9	98.6%
MCL(Ours)	50k	**1455.8**	318.2	60.3	84.8	33.8	70.0	**57.6**	**53.9**	**99.0%**
COINCIDE† [22]	133k	1496.0	298.1	62.9	86.1	32.7	69.2	59.8	55.6	99.3%
ICONS† [54]	133k	1487.1	295.3	63.0	**87.5**	33.0	**70.8**	**60.7**	55.6	99.9%
MCL(Ours)%	133k	**1501.5**	**308.4**	**63.4**	85.7	**33.4**	70.7	60.3	**56.0**	**100.5%**

visual understanding and generalization, we further evaluate on ScienceQA [41], GQA [18], and TextVQA [45], which emphasize reasoning, compositional visual understanding, and text-centric perception, respectively. In all tables and figures, *Rel.* denotes the relative performance to that of the corresponding full-data baseline. For object detection, we report results on the COCO 2017 validation set using Average Precision (AP).

4.2 Main Results and Discussion

Multimodal Data Selection Results. As shown in Tab. 1, experiments on the LLaVA-1.5-mix-665k dataset reveal substantial redundancy in large-scale multimodal instruction data. When retaining only 7.5% of the original training samples, most pruning methods preserve over 90% of the full-data performance, and even random selection avoids a severe performance degradation. This observation indicates that VIT training data contain significant semantic redundancy, leading to inefficient training and motivating the need for principled dataset pruning to improve data quality and training efficiency. Specifically, DataTailor [58] and our MCL, both of which incorporate similarity-aware selection, consistently outperform approaches. This highlights the importance of preserving the information diversity when pruning large-scale VIT datasets. More importantly, although DataTailor jointly considers similarity and representativeness, which also adopts the task-adaptive data allocation strategy, our MCL still achieves stronger overall Rel. performance. We attribute the effectiveness of MCL to the fine-grained, disentangled semantic representation induced by our concept graphs, which enables more precise differentiation among semantic factors between samples. As a result, the retained data better supports generalization across various evaluation tasks. This advantage of our concept graph is further supported by the stability of MCL across benchmarks. Several competing methods exhibit pronounced

Table 2: Comparison to state-of-the-art data pruning methods on COCO 2017 object detection dataset with 70% selected training data.

Method	$AP_{50:95}$	AP_{50}	AP_{75}	AP_{small}	AP_{mid}	AP_{large}
Baseline	40.06	61.12	42.03	19.27	43.39	59.24
Random	38.26 ↓1.80	58.91	39.77	17.83	41.19	56.69
EL2N	34.35 ↓5.71	53.29	36.14	16.58	37.64	48.54
InfoBatch	34.95 ↓5.11	53.86	36.55	17.00	38.74	49.05
DivBS	39.74 ↓0.32	60.84	41.98	18.07	43.14	58.78
PFB	39.69 ↓0.37	60.85	41.74	18.45	43.25	59.07
Ours	40.12 ↑0.06	61.06	42.27	18.95	43.78	59.13

performance imbalance. For instance, InsTag performs strongly on MME but degrades noticeably on SEED-I and GQA, and LESS excels on SEED-I and ScienceQA but underperforms on MME-C. On the contrary, our MCL maintains near-optimal performance across diverse benchmarks. This provides direct evidence of improved generalization rather than overfitting to specific benchmarks.

In addition to performance, our approach offers significant efficiency advantages. As reported in Fig. 2 the entire MCL pruning pipeline including concept graph construction, concept importance estimation, and greedy selection can be completed in 1.7 hours on a 4 * RTX 3090 GPU server. In contrast, methods such as TIVE and DataTailor require expensive pretraining, feature extraction, or clustering procedures. With less than 10% of the overall time cost, our method retains 7.5% of the data but still achieves up to 99% of the full-data performance, demonstrating a substantial improvement in training efficiency.

Object Detection Results. Beyond multimodal instruction tuning, we further evaluate our approach on the purely vision-based object detection task. As reported in Tab. 2, our method achieves higher performance than the full-data baseline while using 70% of the COCO 2017 training data, demonstrating that semantic coverage–aware pruning remains effective even when supervision and task objectives differ substantially from instruction tuning on pure vision data. We also compare our method with general data pruning approaches such as InfoBatch [37], DivBS [16], and PFB [51]. Among them, DivBS and PFB explicitly promote diversity based on embedding representations. Our consistent performance advantage over these methods suggests that fine-grained, concept-level representations provide a more expressive and effective basis for dataset pruning than coarse, coupled vector embeddings. This result further indicates that the benefits of our semantic coverage formulation extend beyond multimodal data and generalize to classical computer vision tasks.

4.3 Ablation Study

Performance at Different Selection Ratios. In Fig. 3, we compare MCL with other recent pruning methods, including COINCIDE [22], D^2-Pruning [32], SemDeDup [1], and Self-Sup [46], across a wide range of selection ratios on LLaVA-1.5-mix-665k. MCL consistently outperforms competing approaches under different data budgets, demonstrating the effectiveness of our framework. In

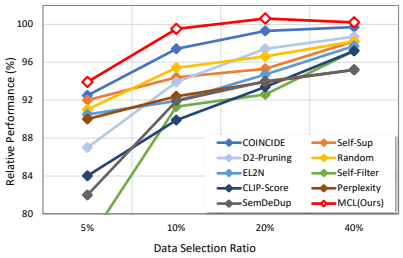


Fig. 3: Relative performance of different data pruning methods at different sampling ratios for the LLaVA-1.5 dataset.

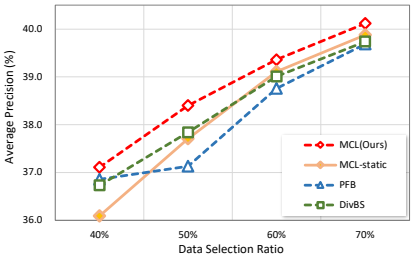


Fig. 4: Different selection ratios on COCO 2017 val. set. Static denotes pruning by ranking the sum of node scores.

particular, MCL surpasses COINCIDE, which estimates density (rarity) using sample-wise embeddings. Similar trends are observed in Fig. 4 on COCO, where MCL based on decoupled concept graphs achieves clear improvements over PFB and DivBS, both of which rely on diversity measured in sample-level embedding space. These results suggest that concept graphs provide a more reliable representation for measuring semantic diversity than compressed sample embeddings.

Static vs. Greedy Selection. We compare our greedy marginal-gain-based pruning with a static alternative that ranks samples solely by their standalone concept importance (Eq. 3), as shown in Fig. 4. On COCO 2017 object detection, the greedy strategy consistently achieves better performance. Static ranking tends to prioritize samples containing rare concepts because importance scores are computed once which can not evolve along with the selection process. Thus, the final selected subset may overemphasize long-tail concepts while underrepresenting dominant concepts that are also essential for learning. This bias becomes increasingly pronounced as the selection ratio decreases, which leads to a larger performance gap between static ranking and the greedy strategy of MCL. These results indicate that effective pruning requires adaptively balancing long-tail concepts with dominant semantics, which is naturally achieved by the our greedy selection based on marginal gains.

Fine-grained Concept Coverage. We investigate whether our method can maintain a high fine-grained semantic coverage at the dataset level. To this end, we compare the concept coverage ratio between the subset selected by MCL and a randomly pruned subset. Quantitative results in Fig. 5 show that MCL covers a substantially larger portion of semantic concepts, confirming that coverage-aware pruning remains effective beyond simple data reduction.

Concept Visualization. To further provide an intuitive understanding of how semantic distribution changes after pruning, we visualize the dataset concept graph before and after pruning. In the Fig. 7, we distinguish the three types of concept nodes using different colors and shapes. Node sizes are proportional to their relative fraction across the dataset, reflecting how often each concept appears in the retained samples. For clarity, we focus on the top 30 most frequent concept nodes in the original dataset graph and the edges among them. Com-

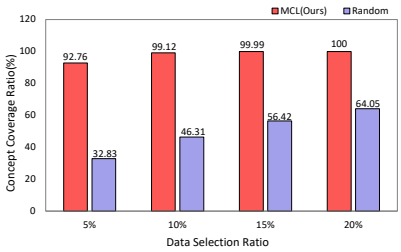


Fig. 5: Semantic concept coverage ratio ($|V_S|/|V_D|$) of MCL and Random pruning on LLaVA-1.5-mix-665k dataset.

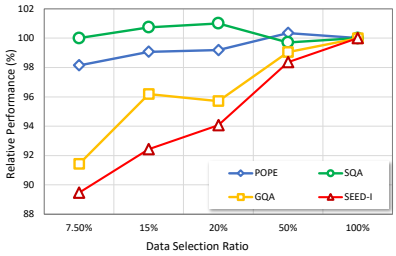
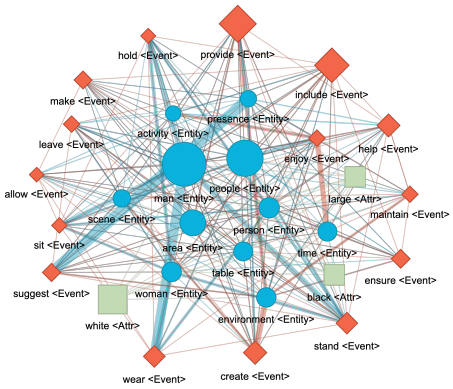
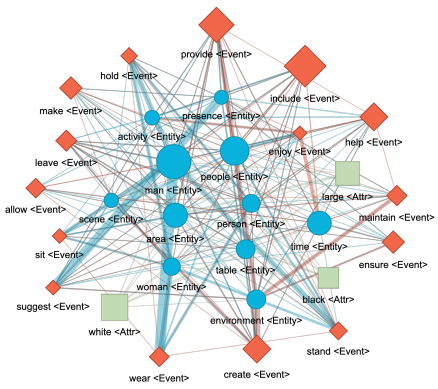


Fig. 6: Relative model performance comparison on different data selection ratios on LLaVA-1.5-mix-665k dataset.



(a) Concept graph before pruning



(b) Concept graph after pruning

Fig. 7: Visualization of the dataset graph (a) before and (b) after pruning. The size of each node is proportional to the fraction of samples containing the concept. Concept distribution becomes more balanced after pruning.

paring the concept graphs before and after pruning reveals a clear redistribution of the semantic concepts. Highly redundant concepts, such as frequent entity nodes (e.g., *man, people*), occupy a smaller proportion after pruning, while less frequent but semantically informative event (e.g., *leave, create*) and attribute nodes gain increased relative prominence. This shift indicates that our method effectively suppresses over-represented concepts while amplifying those under-represented ones, leading to a more balanced semantic distribution. Importantly, this rebalancing is not achieved by discarding dominant concepts entirely, but by moderating their prevalence to make room for rarer concepts. Consequently, the retained subset maintains broad semantic coverage while avoiding excessive redundancy. This observation aligns with our coverage-based objective and provides direct evidence that concept-level pruning can reshape dataset distributions in a principled and interpretable manner.

Effect of Pruning Ratios on Different Benchmarks. Fig. 6 shows that benchmarks exhibit markedly different sensitivities to the data selection ratio.

For POPE and ScienceQA, performance saturates rapidly: with only 7.5% of the data, MCL already reaches or even slightly exceeds the full-data baseline. This indicates that the training set contains substantial redundant data corresponding to these tasks, where a small subset can cover the dominant semantic patterns required for the benchmarks. In contrast, SEED-I and GQA benefit from increased data ratios, reflecting they are more difficult and requires a broader concept distribution. Notably, performance does not increase monotonically with more data on all benchmarks, suggesting that coverage-aware selection can be more effective than retaining larger but redundant subsets.

5 Conclusion

We present MCL, a coverage-aware dataset pruning framework that models large-scale multimodal data through interpretable, fine-grained concept graphs. By decoupling samples into entity, event, and attribute concepts and performing greedy selection based on set-dependent marginal gains, MCL effectively reduces redundancy while preserving semantic diversity. Extensive experiments on both multimodal instruction tuning and object detection demonstrate that MCL achieves strong performance under aggressive pruning, offering substantial improvements in data efficiency with minimal computational overhead.

References

1. Abbas, A.K.M., Tirumala, K., Simig, D., Ganguli, S., Morcos, A.S.: Semdedup: Data-efficient learning at web-scale through semantic deduplication. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. Advances in Neural Information Processing Systems **32** (2019)
3. Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M.L., Paul, M.: Perplexed by perplexity: Perplexity-based data pruning with small reference models. arXiv preprint arXiv:2405.20541 (2024)
4. Bordes, A., Ertekin, S., Weston, J., Bottou, L.: Fast kernel classifiers with online and active learning. Journal of Machine Learning Research **6**(Sep), 1579–1619 (2005)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229 (2020)
6. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
7. Chen, R., Wu, Y., Chen, L., Liu, G., He, Q., Xiong, T., Liu, C., Guo, J., Huang, H.: Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 4156–4172 (2024)

8. Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., Zaharia, M.: Selection via proxy: Efficient data selection for deep learning. In: International Conference on Learning Representations (2019)
9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36**, 49250–49267 (2023)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
11. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. In: International Conference on Learning Representations (2017)
12. Dozat, T., Qi, P., Manning, C.D.: Stanford's graph-based neural dependency parser at the conll 2017 shared task. In: Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. pp. 20–30 (2017)
13. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. Advances in Neural Information Processing Systems **38** (2026)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
15. He, M., Yang, S., Huang, T., Zhao, B.: Large-scale dataset pruning with dynamic uncertainty. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7713–7722 (2024)
16. Hong, F., Lyu, Y., Yao, J., Zhang, Y., Tsang, I., Wang, Y.: Diversified batch selection for training acceleration. In: International Conference on Machine Learning (2024)
17. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al.: Spacy: Industrial-strength natural language processing in python (2020)
18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 6700–6709 (2019)
19. Jiang, A.H., Wong, D.L.K., Zhou, G., Andersen, D.G., Dean, J., Ganger, G.R., Joshi, G., Kaminksy, M., Kozuch, M., Lipton, Z.C., et al.: Accelerating deep learning by focusing on the biggest losers. arXiv preprint arXiv:1910.00762 (2019)
20. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International Conference on Machine Learning. pp. 1885–1894 (2017)
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
22. Lee, J., Li, B., Hwang, S.J.: Concept-skill transferability-based data selection for large vision-language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5060–5080 (2024)
23. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
24. Li, M., Zhang, Y., Li, Z., Chen, J., Chen, L., Cheng, N., Wang, J., Zhou, T., Xiao, J.: From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In: Proceedings of the 2024 Conference of the

North American Chapter of the Association for Computational Linguistics. pp. 7602–7635 (2024)
25. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
27. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
29. Liu, Z., Zhou, K., Zhao, W.X., Gao, D., Li, Y., Wen, J.R.: Less is more: High-value data selection for visual instruction tuning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3712–3721 (2025)
30. Loshchilov, I., Hutter, F.: Online batch selection for faster training of neural networks. arXiv preprint arXiv:1511.06343 (2015)
31. Lu, K., Yuan, H., Yuan, Z., Lin, R., Lin, J., Tan, C., Zhou, C., Zhou, J.: # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In: International Conference on Learning Representations. pp. 36456–36474 (2024)
32. Maharana, A., Yadav, P., Bansal, M.: D2 pruning: Message passing for balancing diversity and difficulty in data pruning. arXiv preprint arXiv:2310.07931 (2023)
33. Mindermann, S., Brauner, J.M., Razzak, M.T., Sharma, M., Kirsch, A., Xu, W., Höltgen, B., Gomez, A.N., Morisot, A., Farquhar, S., et al.: Prioritized training on points that are learnable, worth learning, and not yet learnt. In: International Conference on Machine Learning. pp. 15630–15649 (2022)
34. Mirzasoleiman, B., Cao, K., Leskovec, J.: Coresets for robust training of deep neural networks against noisy labels. Advances in Neural Information Processing Systems **33**, 11465–11477 (2020)
35. Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. Advances in Neural Information Processing Systems **34**, 20596–20607 (2021)
36. Peng, H., Thomson, S., Smith, N.A.: Deep multitask learning for semantic dependency parsing. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 2037–2048 (2017)
37. Qin, Z., Wang, K., Zheng, Z., Gu, J., Peng, X., Zhou, D., Shang, L., Sun, B., Xie, X., You, Y., et al.: Infobatch: Lossless training speed up by unbiased dynamic data pruning. In: International Conference on Learning Representations (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
39. Raju, R.S., Daruwalla, K., Lipasti, M.: Accelerating deep learning with dynamic data pruning. arXiv preprint arXiv:2111.12621 (2021)
40. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)

41. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022)
42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
43. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
44. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. International Conference on Learning Representations (2018)
45. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
46. Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., Morcos, A.: Beyond neural scaling laws: beating power law scaling via data pruning. Advances in Neural Information Processing Systems **35**, 19523–19536 (2022)
47. Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F., Qi, X.: Data pruning via moving-one-sample-out. Advances in Neural Information Processing Systems **36** (2024)
48. Tan, H., Wu, S., Huang, W., Zhao, S., Qi, X.: Data pruning by information maximization. In: International Conference on Learning Representations (2025)
49. Toneva, M., Sordoni, A., des Combes, R.T., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. In: International Conference on Learning Representations (2018)
50. Wei, L., Jiang, Z., Huang, W., Sun, L.: Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. arXiv preprint arXiv:2308.12067 (2023)
51. Wu, D., Guo, Z., Zuo, J., Sang, N., Gao, C.: Partial forward blocking: A novel data pruning paradigm for lossless training acceleration. In: International Conference on Computer Vision. pp. 319–328 (2025)
52. Wu, S., Lu, K., Xu, B., Lin, J., Su, Q., Zhou, C.: Self-evolved diverse data sampling for efficient instruction tuning. arXiv preprint arXiv:2311.08182 (2023)
53. Wu, S., Lu, K., Xu, B., Lin, J., Su, Q., Zhou, C.: Self-evolved diverse data sampling for efficient instruction tuning. arXiv preprint arXiv:2311.08182 (2023)
54. Wu, X., Xia, M., Shao, R., Deng, Z., Koh, P.W., Russakovsky, O.: ICONS: Influence consensus for vision-language data selection. arXiv preprint arXiv:2501.00654 (2024)
55. Xia, M., Malladi, S., Gururangan, S., Arora, S., Chen, D.: Less: selecting influential data for targeted instruction tuning. In: International Conference on Machine Learning. pp. 54104–54132 (2024)
56. Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., Li, P.: Dataset pruning: Reducing training data by examining generalization influence. In: International Conference on Learning Representations (2022)
57. Yang, Y., Mishra, S., Chiang, J., Mirzasoleiman, B.: Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. Advances in Neural Information Processing Systems **37**, 83465–83496 (2024)
58. Yu, Q., Shen, Z., Yue, Z., Wu, Y., Qin, B., Zhang, W., Li, Y., Li, J., Tang, S., Zhuang, Y.: Mastering collaborative multi-modal data selection: A focus on in-

formativeness, uniqueness, and representativeness. In: International Conference on Computer Vision. pp. 155–165 (2025)
59. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
60. Zheng, H., Liu, R., Lai, F., Prakash, A.: Coverage-centric coreset selection for high pruning rates. In: International Conference on Learning Representations (2023)
61. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.: Lima: Less is more for alignment. Advances in Neural Information Processing Systems **36**, 55006–55021 (2023)