PRISM: Self-Pruning Intrinsic Selection Method for Training-Free Multimodal Data Selection

Anonymous ACL submission

Abstract

Visual instruction tuning refines pre-trained Multimodal Large Language Models (MLLMs) to enhance their real-world task performance. However, the rapid expansion of visual instruction datasets introduces significant data redundancy, leading to excessive computational costs. Existing data selection methods predominantly rely on proxy models or loss-based metrics, both of which impose substantial computational overheads due to the necessity of model inference and backpropagation. To address this challenge, we propose **PRISM**, a novel training-free approach for efficient multimodal data selection. Unlike existing methods, PRISM eliminates the reliance on proxy models, warm-up pretraining, and gradient-based optimization. Instead, it leverages Pearson correlation analysis to quantify the intrinsic visual encoding properties of MLLMs, computing a task-specific correlation score to identify high-value instances. This not only enables data-efficient selection, but maintains the model's original performance. Empirical evaluations across multiple MLLMs demonstrate that PRISM reduces the overall time required for visual instruction tuning and data selection to just 30% of conventional methods, while surpassing fully fine-tuned models across eight multimodal and three language understanding benchmarks, achieving a 101.7% relative improvement in final performance.

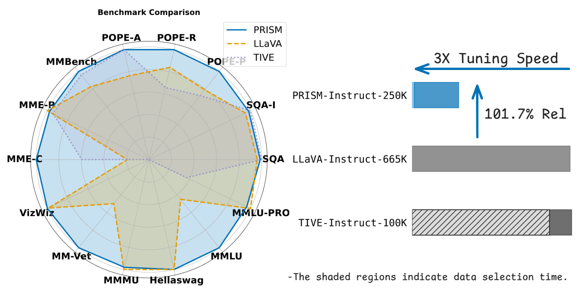

Figure 1: The radar chart illustrates the performance of PRISM, LLaVA, and TIVE across multiple benchmarks. PRISM demonstrates competitive performance while using significantly fewer training samples. The bar chart on the right highlights the data efficiency of PRISM-Instruct-250K, achieving 101.7% relative performance with only 30% of the data used by LLaVA-Instruct-665K and significantly outperforming TIVE-Instruct-100K. The shaded regions indicate data selection time, showing that PRISM achieves 3× faster tuning speed compared to LLaVA and TIVE, emphasizing its efficiency in multimodal instruction tuning.

1 Introduction

The rapid advancement of Multimodal Large Language Models (MLLMs) has significantly transformed artificial intelligence by integrating vision and language processing capabilities (Liu et al., 2024a; Zhu et al., 2023; Dai et al., 2023). Modern MLLMs typically undergo a two-stage training process: (1) large-scale pretraining on web-scale image-text pairs to establish cross-modal alignment, followed by (2) visual instruction tuning on task-specific datasets to enhance instruction-following abilities. While instruction tuning is crucial for achieving strong downstream performance, the exponential growth of low-quality and redundant data (Chen et al., 2024; Wei et al., 2023) in curated datasets poses a major challenge. This proliferation not only increases computational costs but also leads to diminishing returns, highlighting the need for efficient data selection strategies that maximize informativeness while minimizing redundancy.

As training on the full dataset becomes increasingly impractical, selecting the most informative samples is essential for maintaining strong performance while reducing computational overhead. Existing data selection approaches can be broadly classified into two categories: Model-Agnostic Selection and Gradient-Based Selection. Model-Agnostic Selection relies on proxy models, such as pretrained scorers (Chen et al., 2024) or auxiliary MLLMs (Lee et al., 2024), to estimate data importance. However, these methods often introduce bias due to potential misalignment between the proxy and target models. Gradient-Based Selection, on the other hand, utilizes criteria derived from model training dynamics, such as loss-based (Liu et al., 2024d) or influence function-driven metrics (Wu et al., 2025). These approaches are computationally expensive due to the iterative nature of gradient computation. More critically, both paradigms often fail to outperform full-dataset training within practical computational constraints, limiting their real-world applicability.

To address these shortcomings, we introduce PRISM, a novel training-free framework that redefines multimodal data selection by exploiting the intrinsic visual encoding properties of MLLMs. Unlike existing Model-Agnostic and Gradient-Based methods, PRISM represents a third paradigm: Intrinsic Selection. A key challenge in developing such a method is that MLLMs encode rich multimodal interactions in high-dimensional token representations, yet directly leveraging these internal structures for data selection is nontrivial. Unlike Gradient-Based approaches, which capture model learning dynamics, or Model-Agnostic methods, which rely on external scoring heuristics, our Intrinsic Selection extracts meaningful structural information *without* access to model training or auxiliary predictors.

PRISM overcomes this challenge by leveraging the architectural synergy between vision encoders (e.g., CLIP (Radford et al., 2021)) and language models (Li et al., 2023b; Zheng et al., 2023), wherein visual inputs are projected into the LLM's latent space via projectors. Our key insight is that the informational uniqueness of images is inherently captured within the LLM's intermediate token embeddings. By computing pairwise Pearson correlations of token embeddings, PRISM quantifies the representational distinctiveness of visual samples, selecting those that maximize diversity while minimizing redundancy—all without relying on proxy models, gradient computations, or additional training. This method reframes multimodal data selection by leveraging the LLM's intrinsic representations as a quality-sensitive filter, where high-value samples—aligned with the model's semantic priors and exhibiting complementary feature patterns—form unique correlation structures and contribute to increased Shannon entropy, ultimately improving multimodal learning.

We validate PRISM through extensive experiments on a diverse set of multimodal benchmarks, evaluating its efficacy against state-of-the-art data selection methods. Our results demonstrate that MLLMs fine-tuned on PRISM-selected data (`PRISM-Instruct-250K`) outperform models trained on the full dataset while reducing computational costs by 70%. Furthermore, we conduct additional analyses on *Cross-Model Generalization and Scalability* and *Knowledge Retention*, demonstrating that PRISM generalizes effectively across different MLLM architectures and better preserves linguistic capabilities compared to full-dataset training.

Our key contributions are as follows:

- We introduce PRISM, a paradigm shift in multimodal data selection. As the first training-free framework, PRISM fundamentally departs from traditional selection paradigms by eliminating reliance on proxy models, gradient computation, and iterative retraining, offering an efficient yet principled alternative.

- We propose the intrinsic selection mechanism that unlocks the latent structure of multimodal representations. By directly quantifying intrinsic feature redundancy within MLLMs' token embeddings, PRISM enables scalable, high-fidelity data selection—achieving stronger multimodal generalization without additional training overhead.

- Extensive experiments show that PRISM-selected data outperforms full-dataset training while significantly reducing computational costs, making it a practical solution for scalable multimodal learning.

2 Visual Instruction Selection

Visual instruction selection is an approach that can effectively reduce training time of visual instruction tuning by identifying high-value instruction instances. Numerous studies have explored effective methods for selecting such instruction instances while minimizing computational overheads. In this section, we first introduce two fundamental principles for visual instruction selection, which provide a framework for evaluating the effectiveness of different methods in real-world scenarios. Furthermore, we position existing developments in this research area, highlighting key advancements

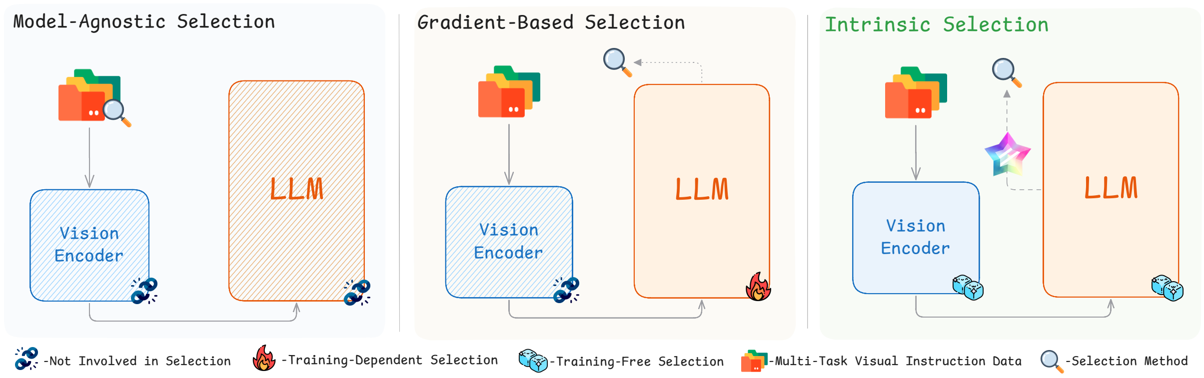

Figure 2: Comparison of data selection paradigms for MLLMs. Model-Agnostic Selection (left) relies on external proxy models without involving the LLM, potentially misaligning with its learned representations. Gradient-Based Selection (middle) uses the LLM's gradients for selection but incurs high computational costs. Intrinsic Selection (PRISM) (right) directly utilizes the LLM's token embeddings, enabling training-free, efficient, and model-aware data selection.

and their implications for optimizing multimodal instruction tuning.

Principle 1: Data selection should not come at the cost of performance. An effective visual instruction selection method should ensure that the model's performance is at least not worse than a fully fine-tuned counterpart. While it is acceptable to achieve better performance with more data, it is not justifiable to compromise model quality in pursuit of dataset reduction.

Principle 2: The time required for visual instruction selection should not exceed the time saved in visual instruction tuning. The primary goal of visual instruction selection is to improve efficiency by reducing the computational burden of instruction tuning. However, if the selection process itself is excessively time-consuming, it defeats the purpose by negating the computational savings gained from dataset reduction. In contrast, an ideal selection method should strike a balance between efficiency and effectiveness, ensuring that the overall training pipeline benefits from reduced resource consumption without introducing additional overhead that outweighs the savings.

As shown in Fig. 2, we categorize current visual instruction selection methods mentioned before into two main types. The first is Model-Agnostic Selection, where the target MLLM remains untouched, and data quality is assessed using a proxy model. Such a proxy model can be a scoring function, such as a pre-trained scoring model (Chen et al., 2024), human reward models, or GPT-based scoring mechanisms (Wei et al., 2023). Some approaches (Lee et al., 2024) also involve training a small MLLM to guide the selection process for the target MLLM. The second category is Gradient-Based Selection, where the target MLLM is first pre-trained on a specific data partition, and data value is subsequently assessed using loss, perplexity, or gradient-based metrics.

Both approaches have inherent limitations:

(1) **Performance degradation**—While the existing two types of selection methods effectively filter a subset of the data, they often degrade the model performance compared to full fine-tuning (see in Table 1). This contradicts *Principle 1*, as the objective is to construct a stronger model with fewer data, rather than a weaker model due to reduced data availability.

(2) **High computational cost**—Gradient-based selection is computationally prohibitive. For instance, TIVE (Liu et al., 2024d) employs LoRA-based warm-up training on the target MLLM before computing gradient vectors. However, the time required for this process often surpasses the time saved in instruction tuning (see in Table 4), violating *Principle 2* and rendering it impractical.

(3) **Proxy model bias**—To mitigate computational overhead, some methods rely on proxy models, keeping the target MLLM independent during training. However, this introduces bias from the pre-trained proxy model (e.g., GPT or a human reward model) or a warm-up trained small MLLM, which may not generalize well to the target MLLM. Since proxy

models and warm-up data significantly influence selection results, a general selection strategy that excludes the target MLLM will yield the same selected data across different MLLMs, despite their distinct data requirements. Consequently, such approaches fail to provide optimal data selection tailored to specific MLLMs.

To overcome these challenges, we introduce PRISM, a self-PRuning Intrinsic Selection Method for training-free multimodal data selection.

3 PRISM

The PRISM framework establishes a new paradigm for multimodal data selection by directly harnessing the intrinsic representation structures of MLLMs. Unlike existing methods that depend on external heuristics or model-driven proxies, PRISM leverages the model's intrinsic encoding mechanisms to assess data informativeness. Modern MLLMs, such as LLaVA (Liu et al., 2024a), unify visual and textual modalities through a vision encoder and projector, embedding images into the LLM's latent space—where their uniqueness is inherently captured. This approach ultimately enhances performance while reducing training time by 70%. Our initial research (as in Fig. 3) revealed that layer-wise token embeddings inherently capture structural distinctions between informative and redundant samples. Inspired by this, we explored the statistical dependencies within these embeddings to systematically identify high-value data instances. These findings ultimately led to the design of PRISM, a method that selects informative samples without relying on external supervision (e.g., proxy models or gradient-based computations).
PRISM formalizes this approach in a three-stage pipeline: feature representation, correlation analysis, and self-pruning selection. As we will see in the performance evaluation, PRISM offers a scalable and computationally efficient solution to multimodal data selection.

3.1 Feature Representation and Correlation Analysis

Let $\mathcal{D} = \{I_1, I_2, \dots, I_N\}$ denote the image dataset for target task $\mathcal{T}$. For each image I_i, the vision encoder (VE) extracts and projects visual embeddings into the LLM's latent space:

$$v_i = \mathrm{VE}(I_i) \in \mathbb{R}^{d_v}, \quad z_i = \mathrm{Proj}(v_i) \in \mathbb{R}^d \quad (1)$$

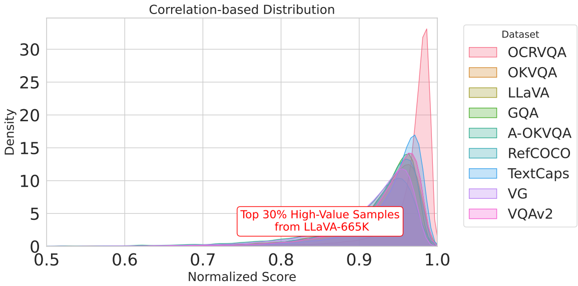

Figure 3: Correlation-based distribution of multimodal data across datasets. The selection strategy prioritizes samples that balance redundancy reduction and information diversity, ensuring that high-correlation images do not dominate while preserving a broad range of semantic variance. The highlighted region represents the top 30% of high-value samples identified from LLaVA-665K.

where $\mathrm{Proj} : \mathbb{R}^{d_v} \to \mathbb{R}^d$ is a linear projector. The LLM processes z_i through transformer layers, with averaged token features from layer l computed as:

$$F_i = \frac{1}{T} \sum_{t=1}^{T} \mathrm{LLM}^{(l)}(z_i)_t \in \mathbb{R}^d \quad (2)$$

where T denotes the number of tokens. We hypothesize that images with divergent feature correlations provide complementary information. This is quantified through Pearson analysis:

$$P_{ij} = \frac{\mathbb{E}[(F_i - \mu_i)(F_j - \mu_j)]}{\sigma_i \sigma_j}, \quad C_i = \sum\nolimits_{j=1}^{N} P_{ij} \quad (3)$$

where μ_i, σ_i are mean and standard deviation of F_i, and C_i measures alignment with $\mathcal{D}$'s feature distribution.

3.2 Self-Pruning Selection

Images with the lowest C_i values (i.e., those in the bottom τ% of the sorted correlation scores) are selected as high-value candidates. This selection strategy is guided by three factors:
Reduction of Feature Redundancy: High correlation images ($\uparrow C_i$) exhibit substantial semantic overlap, contributing diminishing returns during training.
Information-Theoretic Diversity: Low correlation samples ($\downarrow C_i$) maximize the Shannon entropy of the selected subset, as formally analyzed in Appendix C.
Outlier Resilience: Unlike variance-based selection, Pearson correlation's scale invariance ensures robustness to embedding magnitude variations caused by projector miscalibrations.

Formally, given a threshold τ (e.g., $\tau = 30\%$), the selected subset is defined as:

$$\mathcal{D}_{\text{selected}} = \{I_i \mid C_i \leq Q_\tau(C)\}, \tag{4}$$

where Q_τ denotes the τ-th percentile of correlation scores. Algorithm 1 is organized into three main phases, corresponding to the selection strategy's core factors:

Intrinsic Feature Extraction (Step 1): This phase computes visual embeddings and projects them into the LLM's latent space, obtaining layer-wise averaged token features. By capturing intrinsic semantic information, it prepares the feature space for correlation analysis, laying the foundation for *Outlier Resilience*.

Correlation Analysis (Step 2): The Pearson correlation matrix is computed to evaluate feature similarity across all images. Summing each row gives the total correlation score for each image, C_i, which quantifies its semantic redundancy. This phase directly targets the *Reduction of Feature Redundancy* by identifying images with high semantic overlap.

Self-Pruning Selection (Step 3): By sorting and selecting images with the lowest C_i scores, this step maximizes diversity while avoiding redundant samples. It achieves *Information-Theoretic Diversity* by preserving images that contribute the most to the subset's entropy, ensuring data efficiency.

Algorithm 1 PRISM Data Selection

Require: Image dataset $D = \{I_1, \dots, I_N\}$, vision encoder VE, projector Proj, LLM layer l, threshold τ

Ensure: Selected subset D_{selected}

- **Step 1: Intrinsic Feature Extraction**
- **for** each image $I_i \in D$ **do**
  - Compute visual embedding: $v_i \leftarrow \text{VE}(I_i)$
  - Project to LLM space: $z_i \leftarrow \text{Proj}(v_i)$
  - Extract layer-l features:

$$F_i \leftarrow \frac{1}{T} \sum_{t=1}^{T} \text{LLM}^{(l)}(z_i)_t$$

- **end for**
- **Step 2: Correlation Analysis**
- Construct feature matrix $F \in \mathbb{R}^{N \times d}$
- Compute Pearson matrix $P_{ij} \leftarrow \frac{\text{cov}(F_i, F_j)}{\sigma_{F_i} \sigma_{F_j}}$
- Score images: $C_i \leftarrow \sum_{j=1}^{N} P_{ij}, \forall i$
- **Step 3: Self-Pruning Selection**
- Sort indices: $\text{argsort}(C) \leftarrow [i_1, \dots, i_N]$ s.t. $C_{i_1} \leq \cdots \leq C_{i_N}$
- Select subset: $D_{\text{selected}} \leftarrow \{I_{i_k} \mid k \leq \lfloor \tau N \rfloor\}$
- **return** D_{selected}

4 Experiments

We first present our experimental setup and evaluation benchmarks, followed by comparisons with state-of-the-art methods. Next, we analyze our method's behavior and effectiveness across various dimensions. Additionally, we evaluate the transferability of our strategy to unseen tasks and model architectures. Finally, we conduct ablation studies to assess the contribution of each component.

4.1 Experiment Setup

Dataset & Model: We evaluate PRISM on the visual instruction tuning dataset LLaVA-665K (Liu et al., 2024a), using LLaVA-1.5-7B (Liu et al., 2024a) as our primary base model. All experiments are conducted for one epoch following the official fine-tuning hyperparameters. To ensure a fair comparison, we maintain a consistent training environment across all evaluations.

Baselines: We compare PRISM against a comprehensive set of data selection baselines, including Random Selection, Instruction Length, Perplexity (Liu et al., 2024d), GraNd (Paul et al., 2023), EL2N (Paul et al., 2023), InstructionGPT-4 (Wei et al., 2023), SELF-FILTER (Chen et al., 2024), TIVE (Liu et al., 2024d), COINCIDE (Lee et al., 2024), DataTailor (Yu et al., 2024a), and ICONS (Wu et al., 2025). To ensure fair comparisons, we adopt the experimental settings and incorporate results from ICONS (Wu et al., 2025) and TIVE (Liu et al., 2024d).

Benchmarks: Following the evaluation framework of LLaVA-1.5 (Liu et al., 2024a), we assess the effectiveness of PRISM across a diverse set of multimodal benchmarks designed to test various capabilities of MLLMs. These benchmarks are grouped into three main categories: understanding and reasoning, factual consistency and generalization, and visual conversation and core multimodal skills.

For understanding and reasoning, we evaluate the model's ability to perform multiple-choice tasks (MMBench (Liu et al., 2024c)), scientific question answering (ScienceQA (Lu et al., 2022)), and multimodal reasoning (MME (Yin et al., 2023)). For factual consistency and generalization, we measure the model's tendency for hallucination (POPE (Li et al., 2023a)) and its zero-shot generalization ability on unseen visual queries (VizWiz (Gurari et al., 2018)). Finally, for visual conversation and core multimodal skills, we assess the model's conversational capabilities (MM-Vet (Yu et al., 2024b)) and its proficiency in perception, knowledge integration, and reasoning (MMMU (Yue et al., 2024)).

4.2 Main Results

We present a comprehensive evaluation of PRISM across multiple settings. First, in Table 1, we compare PRISM with the baseline methods selected above on the LLaVA-1.5-7B model (Liu et al., 2024a). This demonstrates its superior performance in multimodal data selection. Next, Table 2 showcases the results of PRISM across different MLLMs, which highlights its generalizability and robustness. Finally, Table 3 focuses on PRISM's text-only capabilities, which provides insights into its effectiveness in uni-modal settings.

We further analyze these results as follows:

Superior Multimodal Understanding. As shown in Table 1, PRISM achieves the best performance across 11 multimodal benchmarks, surpassing full-dataset fine-tuning by **1.7%** in relative performance. Notably, PRISM excels in instruction-sensitive tasks: it outperforms full fine-tuning on MMBench (65.2 vs. 64.3) and MM-Vet (32.0 vs. 31.1), demonstrating its ability to select samples that enhance complex reasoning and visual conversation capabilities. The improvements are particularly significant compared to gradient-based methods like GraNd (62.9 vs. 65.2 on MMBench), highlighting the limitations of loss-driven selection in multimodal contexts.

Hallucination Mitigation. PRISM achieves the highest scores on all POPE subsets (87.7/88.7/85.5), outperforming even specialized hallucination reduction methods like ICONS (87.5). This suggests that low-correlation samples inherently reduce the model's tendency to generate inconsistent facts, as they avoid overfitting to spurious text-visual correlations prevalent in redundant data.

Balanced Efficiency and Performance. While gradient-based methods like TIVE achieve comparable average performance (100.6% rel.), their total time costs (selection + training) often exceed full fine-tuning due to iterative model updates. In contrast, PRISM achieves higher accuracy (**101.7%**) and reduces total time by **70%**. This efficiency stems from its training-free advantage: feature extraction and correlation computation are executed in a single forward pass and offline batched processing, respectively, with negligible overhead compared to full training cycles. Remarkably, PRISM simultaneously enhances spatial reasoning capabilities (**330.0** on MME-C vs. 311.9 for full fine-tuning), validating that its selection criteria preserve geometrically informative samples often lost in random or length-based pruning.

4.3 Model Behavior Analysis

Cross-Model Generalization and Scalability. PRISM is designed to identify high-value data that remains effective across different model architectures and scales. To validate this, we assess whether data selected using one model setup can benefit others. While our subset was initially selected with LLaVA-1.5-7B, we further evaluate its effectiveness on two additional model configurations. The detailed architectures of these models are summarized in Appendix 7. The results approve that PRISM captures generally useful training samples rather than those tailored to a specific model.

As shown in Table 2, PRISM demonstrates strong cross-architecture and cross-scale generalization capabilities. The subset selected using the 7B model achieves competitive performance across different model sizes and architectures, suggesting that our method captures fundamental visual-language understanding capabilities that are transferable and scalable. This highlights the robustness of PRISM in identifying high-value data points that generalize well across diverse multimodal model configurations.

Language Knowledge Retention. While visual instruction tuning significantly enhances performance on vision-centric tasks, it often leads to a degradation in the model's ability to handle text-only tasks (Zhang et al., 2024). To assess the text-only performance, we evaluate PRISM on a range of benchmarks accordingly, including interdisciplinary knowledge assessments such as MMLU (Hendrycks et al., 2021) and MMLU-PRO (Wang et al., 2024), as well as reasoning tasks like HellaSwag (Zellers et al., 2019). These benchmarks are designed to test the model's ability to retain and utilize its original language understanding capabilities after multimodal fine-tuning.

The results in Table 3 further demonstrate that PRISM in some cases can even improve the model's performance on text-only tasks, suggesting that our data selection method effectively mitigates

Method	SQA	SQA-I	VizWiz	POPE-P	POPE-R	POPE-A	MM-Vet	MMBench	MME-P	MME-C	MMMU	Rel. (%)
Full-Finetune	69.4	66.8	50.0	86.1	87.3	84.2	31.1	64.3	1510.7	311.9	35.4	100%
Random	65.5	64.5	48.1	85.1	84.6	83.6	30.2	55.5	1492.0	233.5	30.5	93.2%
Length	66.8	66.7	47.0	85.4	85.5	84.1	31.5	57.0	1422.1	306.0	33.1	96.6%
EL2N	70.2	70.6	44.4	85.6	85.6	85.6	-	61.6	1356.5	294.7	-	97.2%
Perplexity	70.5	67.9	-	83.3	83.3	83.3	-	62.3	1393.3	260.7	-	95.8%
GraNd	71.4	68.4	37.8	82.5	82.5	82.5	-	62.9	1400.5	287.1	-	94.6%
TIVE	72.2	70.6	-	85.6	85.6	85.6	-	63.2	1433.0	322.1	-	100.6%
InstructionGPT-4	-	-	-	-	-	-	-	31.4	463.3	-	-	39.75%
Self-Filter	-	61.4	53.2	83.8	83.8	83.8	-	61.4	1306.2	-	-	96.1%
COINCIDE	-	69.2	46.8	86.1	86.1	86.1	-	63.1	1495.6	-	-	99.3%
ICONS	-	70.8	-	87.5	87.5	87.5	-	63.1	1485.7	-	-	101.0%
DataTailor	71.0	-	49.5	85.3	85.3	85.3	-	-	1476.1	319.2	-	99.9%
PRISM (Ours)	**71.3**	**69.1**	**50.1**	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	1470.0	**330.0**	34.7	**101.7%**

Table 1: Evaluation of PRISM against full fine-tuning and existing data selection approaches across multiple multimodal understanding benchmarks. PRISM achieves superior performance, surpassing full fine-tuning while significantly reducing computational costs. Metrics in **bold** indicate improvements over the full fine-tuning baseline. For POPE, we report the average score across three subsets for certain baselines due to the unavailability of complete results.

Model	SQA	SQA-I	VizWiz	POPE-P	POPE-R	POPE-A	MM-Vet	MMBench	MME-P	MME-C	MMMU	Rel. (%)
LLaVA-Phi2-3B	75.3	72.7	41.2	87.3	88.6	86.1	35.6	68.7	1467.7	298.0	37.7	100%
PRISM-3B	**76.3**	**72.8**	40.9	**87.5**	**88.8**	**86.5**	34.1	**68.9**	**1485.5**	**305.0**	37.6	**100.1%**
LLaVA-Vicuna-7B	69.4	66.8	50.0	86.1	87.3	84.2	31.1	64.3	1510.7	311.9	35.4	100%
PRISM-7B	**71.3**	**69.1**	**50.1**	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	1470.0	**330.0**	34.7	**101.7%**
LLaVA-Vicuna-13B	74.4	71.6	53.6	87.4	88.0	85.6	36.1	67.7	1531.3	295.4	35.1	100%
PRISM-13B	**74.5**	**71.8**	53.1	**87.7**	**88.4**	**85.7**	**36.4**	65.8	**1538.5**	**307.5**	**35.7**	**100.4%**

Table 2: Performance on Cross-Model Generalization and Scalability with PRISM.

Model	Hellaswag	MMLU	MMLU-PRO	Rel. (%)
LLaVA-Phi2-3B	66.0	50.5	9.1	100%
PRISM-3B	**67.4**	**52.7**	8.6	**100.3%**
LLaVA-Vicuna-7B	66.5	35.0	17.8	100%
PRISM-7B	**66.5**	**41.1**	15.7	**101.9%**
LLaVA-Vicuna-13B	69.5	36.2	6.8	100%
PRISM-13B	**69.6**	**39.5**	**12.4**	**130.6%**

Table 3: Results on language benchmarks.

Method	Data Selection	Visual Instruction Tuning	Overall
Full-Finetune	-	94 (Hours)	94
TIVE	87 (Hours)	14 (Hours)	101 (+7.5%)
PRISM	1.5 (Hours)	28 (Hours)	29.5 (-71%)

Table 4: Wall-clock runtime (measured as A100 80G GPU) for total computation cost.

the knowledge forgetting problem commonly associated with visual instruction tuning. This highlights the dual benefit of PRISM: enhancing multimodal task performance while maintaining the model's foundational language capabilities.

4.4 Ablation Study

To validate the design choices of PRISM, we conduct systematic ablation studies on three key components: LLM layer selection, correlation-based scoring, and token aggregation for image representation.

Influence of LLM Layer Selection. We first investigate how different transformer layers impact PRISM's performance by extracting features from three representative layers: *Shallow Layer 1*, which captures low-level visual patterns such as edges and textures; *Middle Layer 16*, which balances visual and semantic features; and *Deep Layer 32*, which encodes high-level semantic abstractions. As shown in Table 5, PRISM achieves the highest performance when using shallow layer features, outperforming deeper layers by 2.8%. This result indicates that early-layer embeddings sufficiently capture the necessary information for redundancy detection, while deeper layers may overfit to task-specific semantics, leading to reduced generalizability.

Impact of Correlation-based Selection. We evaluate PRISM's correlation-based selection strategy by partitioning the dataset into three groups based on their Pearson correlation scores: *Low-Correlation Group*, *Medium-Correlation Group*, and *High-Correlation Group*. The results in Table 5 demonstrate that selecting low-correlation samples leads to the highest performance, outperforming the high-correlation group by 3.7%. This supports

Method	SQA	SQA-I	VizWiz	POPE-P	POPE-R	POPE-A	MM-Vet	MMBench	MME-P	MME-C	MMMU	Rel. (%)
Deep Layer	71.2	69.1	51.6	86.6	88.0	84.2	31.1	62.9	1477.0	254.0	34.5	97.2%
Middle Layer	70.9	69.1	47.7	86.5	87.8	84.2	31.9	65.0	1517.1	276.0	34.9	97.9%
✓ Shallow Layer	**71.3**	**69.1**	50.1	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	1470.0	**330.0**	34.7	**100.0%**
High Correlation	70.6	68.0	48.1	85.8	87.6	83.9	30.7	64.0	1428.5	275.3	33.5	96.3%
Moderate Correlation	71.0	69.7	48.3	85.9	86.7	84.0	30.0	64.2	1509.0	286.0	34.1	97.3%
✓ Low Correlation	**71.3**	69.1	**50.1**	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	1470.0	**330.0**	**34.7**	**100.0%**
Last Token	69.9	67.3	49.4	87.4	88.3	85.0	31.6	62.6	1471.0	272.0	35.3	97.4%
✓ Avg Pooling	**71.3**	**69.1**	**50.1**	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	1470.0	**330.0**	34.7	**100.0%**

Table 5: Ablation study results on LLM layer selection, correlation-based scoring, and token aggregation for image representation.

our hypothesis that prioritizing samples with minimal correlation maximizes informational diversity, whereas high-correlation samples tend to be redundant, diminishing their contribution to multimodal learning. Our findings underscore the effectiveness of leveraging feature correlation as a criterion for efficient data selection (see Appendix C for further analysis).

Effect of Token Aggregation Strategy. Finally, we examine how different token aggregation methods influence image feature modeling. We compare two approaches: *Average Token*, which computes a global average over all transformer tokens, and *Last Image Token*, which uses only the final image token in the sequence. As shown in Table 5, the average token method achieves the best performance, surpassing the last image token by 2.6%. This result aligns with PRISM's design principle that averaging token representations captures holistic visual semantics, whereas relying on the last token may introduce positional biases or task-specific artifacts. These findings validate our choice of average pooling as a more robust and generalizable strategy for training-free multimodal data selection.

5 Related Work

Visual Instruction Tuning: Visual instruction tuning is essential for aligning MLLMs with both practical applications and academic benchmarks. Early methods relied on synthetic visual instructions, which performed well in conversations but struggled on rigorous benchmarks. A hybrid approach later emerged, integrating synthetic data with academic datasets to improve training diversity. This advancement has enhanced models like LLaVA (Liu et al., 2024b), InstructBLIP (Dai et al., 2023), and Cambrian (Tong et al., 2024), enabling better visual-linguistic understanding. Beyond task performance, visual instruction tuning improves model alignment with user expectations, ensuring both practical utility and strong academic performance.

Visual Instruction Selection: Despite the strong performance of MLLMs, the rapid growth of visual instruction datasets has introduced significant redundancy, similar to challenges in LLMs (Zhou et al., 2024; Chen et al., 2023; Xia et al., 2024). State-of-the-art models like BLIP3 (Xue et al., 2024), InternVL2.5 (Chen et al., 2025), and LLaVA-OneVision (Li et al., 2024) rely on billions of instructions to enhance understanding, but their massive scale leads to substantial computational costs, often requiring hundreds to thousands of GPU hours.

To address this, various data selection strategies aim to reduce redundancy while preserving performance. TIVE (Liu et al., 2024d) selects valuable data based on gradient similarity but requires additional training on downstream tasks. SELF-FILTER (Chen et al., 2024) uses an auxiliary evaluation model to prioritize high-value samples. COINCIDE (Lee et al., 2024) clusters data by conceptual and skill-based representations, while InstructionGPT-4 (Wei et al., 2023) filters 200 instructions for MiniGPT-4 (Zhu et al., 2023), though it lacks scalability. ICONS (Wu et al., 2025) extends LESS (Xia et al., 2024) by incorporating specialist influence estimation for instruction tuning. DataTailor (Yu et al., 2024a) selects data based on informativeness, uniqueness, and representativeness to retain the most relevant samples.

6 Conclusion

PRISM leverages MLLMs' intrinsic cross-modal alignment to select high-value samples using Pearson correlations of token embeddings, requiring no proxy models or training. It achieves 70% cost reduction while maintaining performance, setting a new standard for efficient multimodal learning.

Limitations

A limitation of this work is the static nature of the data selection strategy, which only handles text and image modalities. Extending this approach to include video and sound could introduce challenges due to the temporal and sequential properties of these modalities.

Additionally, our method does not incorporate dynamic data selection during training. Adapting the selection process over time could improve model efficiency by focusing on the most relevant data at each stage, particularly for large and diverse datasets.

References

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. Llava steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *Preprint*, arXiv:2412.12359.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. 2023.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. *Preprint*, arXiv:2406.10995.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024d. Less is more: High-value data selection for visual instruction tuning. *Preprint*, arXiv:2403.09559.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2023. Deep learning on a data diet: Finding important examples early in training. *Preprint*, arXiv:2107.07075.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

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

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2025. Icons: Influence consensus for vision-language data selection. *Preprint*, arXiv:2501.00654.

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

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2024a. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. *arXiv preprint arXiv:2412.06293*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal llms without text-only forgetting. *Preprint*, arXiv:2406.03496.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

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

A Dataset Details

We present the detailed composition of the PRISM-Instruct-250K dataset, which spans multiple visual question answering (VQA), image understanding, and text-based tasks. This diverse selection ensures a comprehensive representation of multimodal learning challenges. The table below shows the distribution of samples across different data sources.

Dataset	Number of Samples
LLaVA	53,591
VQAv2	27,567
OKVQA	2,997
A-OKVQA	22,032
RefCOCO	16,933
VG	28,777
GQA	24,023
OCRVQA	26,638
TextCaps	7,311
Text-Only	40,688
Total	250,557

Table 6: Sample distribution of the PRISM-selected Instruct-250K dataset.

B Model Architectures

In our experiments, we assess PRISM's transferability across various model architectures and scales, following the methodology outlined in (Bi et al., 2025). The models tested include LLaVA-Vicuna-7B, LLaVA-Phi2-3B, and LLaVA-Vicuna-13B. Each model consists of a vision encoder, a projector, and a language model. The table below summarizes their configurations.

Table 7: Architectural configurations of the models used in our experiments.

Model	Vision Encoder	Language Model
LLaVA-Vicuna-7B	CLIP ViT-L/14 336px	Vicuna v1.5 7B
LLaVA-Phi2-3B	SigLIP-SO400M-Patch14-384	Phi-2 2.7B
LLaVA-Vicuna-13B	CLIP ViT-L/14 336px	Vicuna v1.5 13B

C Shannon Entropy and Feature Diversity

Let $F = \{f_1, f_2, \ldots, f_N\}$ denote the set of feature vectors extracted from the dataset D, where each $f_i \in \mathbb{R}^d$ corresponds to the averaged token embedding of image I_i. The Shannon entropy $H(F)$ of the feature set is defined as:

$$H(F) = -\sum_{i=1}^{N} p(f_i) \log p(f_i), \tag{5}$$

where $p(f_i)$ is the probability density of feature f_i. In practice, $p(f_i)$ can be approximated using kernel density estimation or other non-parametric methods. However, directly maximizing $H(F)$ is computationally infeasible for large datasets. Instead, we use the Pearson correlation matrix P as a proxy for feature diversity.

The Pearson correlation matrix P captures pairwise linear dependencies between feature vectors. For a given feature f_i, the correlation score is given by:

$$C_i = \sum_{j=1}^{N} P_{ij}, \tag{6}$$

which quantifies its overall alignment with the dataset. Low C_i values indicate that f_i is weakly correlated with other features, suggesting that it contributes unique information to the dataset.

Let $F_{\text{selected}} = \{f_i \mid C_i \leq Q_\tau(\{C_j\}_{j=1}^N)\}$ denote the subset of features selected by PRISM, where Q_τ is the τ-th percentile of correlation scores. The entropy of F_{selected} can be approximated as:

$$H(F_{\text{selected}}) \approx -\sum_{f_i \in F_{\text{selected}}} p(f_i) \log p(f_i). \tag{7}$$

By selecting features with minimal C_i, we implicitly minimize the pairwise dependencies within F_{selected}, thereby maximizing the entropy $H(F_{\text{selected}})$. This is because low-correlation features are less likely to share redundant information, leading to a more diverse and informative subset.

We now formalize this intuition with the following theorem:

Theorem 1 (Entropy Maximization via Low-Correlation Selection): Let F be a set of feature vectors with correlation matrix P. For any subset $F_{\text{selected}} \subseteq F$, the Shannon entropy $H(F_{\text{selected}})$ is maximized when F_{selected} consists of features with minimal row-wise sums:

$$C_i = \sum_{j=1}^{N} P_{ij}. \tag{8}$$

Proof: The proof follows from the properties of Shannon entropy and the definition of Pearson correlation. Let $F_{\text{selected}} = \{f_i \mid C_i \leq Q_\tau(\{C_j\}_{j=1}^N)\}$. By construction, F_{selected} contains features that are minimally correlated with the rest of the dataset. This implies that the pairwise dependencies within F_{selected} are reduced, leading to a higher entropy $H(F_{\text{selected}})$.

Formally, for any two subsets F_1 and F_2 with $H(F_1) > H(F_2)$, the features in F_1 exhibit lower pairwise correlations on average. Thus, selecting features with minimal C_i ensures that $H(F_{\text{selected}})$ is increased.

The above theorem provides a theoretical foundation for PRISM's low-correlation selection strategy. By prioritizing features with minimal C_i, PRISM ensures that the selected subset F_{selected} is both diverse and informative, aligning with the principles of information-theoretic feature selection.