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

Anonymous ACL submission

Abstract

Visual instruction tuning adapts pre-trained Multimodal Large Language Models (MLLMs) to follow human instructions for real-world applications. However, the rapid growth of these datasets introduces significant redundancy, leading to increased computational costs. Existing methods for selecting instruction data aim to prune this redundancy, but predominantly rely on computationally demanding techniques such as proxy-based inference or training-based metrics. Consequently, the substantial computational costs incurred by these selection processes often exacerbate the very efficiency bottlenecks they are intended to resolve, posing a significant challenge to the scalable and effective tuning of MLLMs. To address this challenge, we first identify a critical, yet previously overlooked, factor: the anisotropy inherent in visual feature distributions. We find that this anisotropy induces a *Global Semantic Drift*, and overlooking this phenomenon is a key factor limiting the efficiency of current data selection methods. Motivated by this insight, we devise **PRISM**, a training-free framework for efficient visual instruction selection. PRISM effectively removes the corrupting influence of global background features by modeling the intrinsic visual semantics via implicit re-centering. Empirically, PRISM reduces the end-to-end time for data selection and model tuning to just 30% of conventional pipelines. More remarkably, it achieves this efficiency while simultaneously enhancing performance, surpassing models fine-tuned on the full dataset across various multimodal and language understanding benchmarks, culminating in a 101.7% relative performance of the baseline. The code is available for access via this repository.

1 Introduction

The advent of MLLMs marks a significant milestone in artificial intelligence, endowing models with remarkable capabilities to process and reason over interleaved visual and textual information

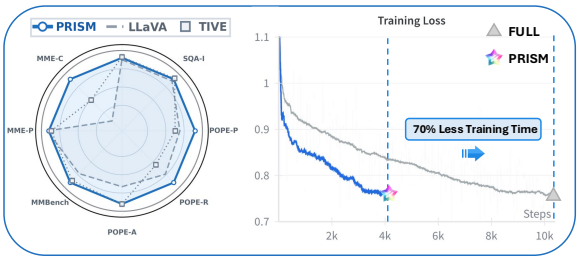

Figure 1: **PRISM achieves state-of-the-art performance with substantially greater training efficiency. (Left)** On a suite of multimodal and language benchmarks, PRISM uniformly outperforms strong baselines like TIVE. **(Right)** The training loss curve demonstrates that PRISM converges faster, reducing training time by 70% compared to the LLaVA baseline.

(Liu et al., 2024a; Zhu et al., 2023; Dai et al., 2023). The standard paradigm for building these models is a two-stage process: (1) large-scale pre-training on web-scraped image-text pairs to establish foundational cross-modal alignment, followed by (2) visual instruction tuning to hone the model's ability to follow complex, task-oriented instructions. While instruction tuning is critical for unlocking state-of-the-art performance, its effectiveness is increasingly challenged by the data it relies on. Publicly curated datasets are expanding exponentially but are often saturated with low-quality and redundant examples (Chen et al., 2024a; Wei et al., 2023). Naively training on this noisy data inflates computational costs and, paradoxically, can degrade final model performance. This predicament has created an urgent need for efficient visual instruction selection methods that can identify a compact, high-quality subset of data.

Existing selection strategies predominantly fall into two categories. The first, **Proxy-Based Selection**, employs external models such as specialized scorers (Chen et al., 2024a) or auxiliary MLLMs (Lee et al., 2024) to estimate the importance of each data point. This approach, however, introduces substantial computational overhead from the

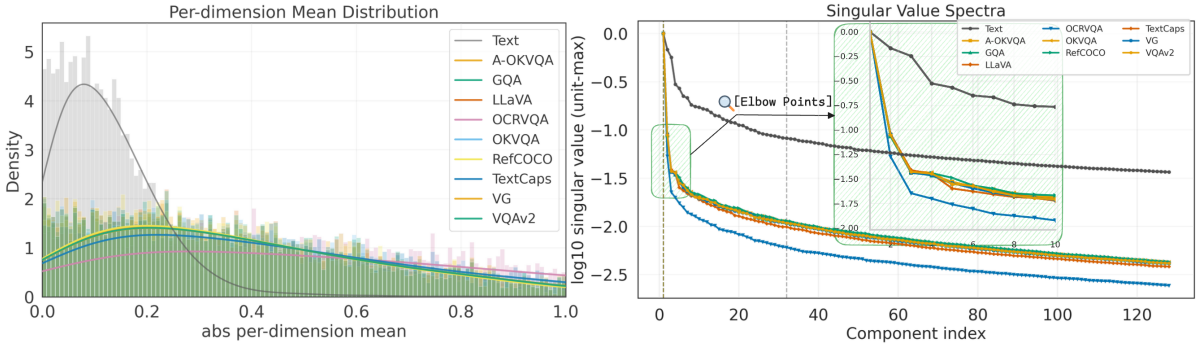

Figure 2: **Visual Diagnosis of Representation Anisotropy in Pre-trained MLLM Features. (Left)** The per-dimension mean distribution reveals a fundamental disparity: unlike text features (grey) which are well-centered around zero, all visual instruction datasets exhibit a significant non-zero mean. This provides direct evidence that visual embeddings occupy a biased, narrow cone, which is the geometric origin of *Global Semantic Drift*. **(Right)** The singular value scree plot confirms this diagnosis. The sharp "elbow point" indicates that the feature variance is confined to a few dominant dimensions—a classic symptom of representation degeneration. Detailed statistics and qualitative case studies are provided in Appendix H and Fig. 9.

proxy model's training and inference demands and is susceptible to selection bias when its objectives or architecture misalign with the target model. The second category, **Training-Based Selection**, leverages signals from the target model's own training dynamics, using metrics based on training loss (Liu et al., 2024d) or influence functions (Wu et al., 2025). While this avoids proxy misalignment, these methods are often prohibitively expensive, requiring iterative gradient computations that can be as costly as training on the full dataset itself.

A critical limitation unites both paradigms: their substantial computational costs often negate the intended efficiency gains and can even lead to performance degradation under practical resource constraints. This reveals a fundamental gap in the current landscape—the lack of a selection method that is both effective at identifying high-quality data and genuinely efficient to deploy. In this work, we argue that the key to bridging this gap lies not in designing more complex selection architectures, but in diagnosing and correcting a fundamental flaw within the MLLM's own visual representations. Our investigation begins with an analysis of the pre-trained visual feature distribution, where we uncover a critical, yet previously overlooked, property: *representation anisotropy*.

As illustrated in Figure 2, we first observe that unlike text features, the mean embedding vector of visual features across diverse datasets is consistently non-zero. This signifies that the representations are confined to a narrow cone in high-dimensional space, a geometric bias that systematically distorts common similarity metrics (Levi and Gilboa, 2024). This distortion gives rise to a phenomenon we term *Global Semantic Drift*, where global background context from different visual domains misleadingly inflates cosine similarity scores. A deeper analysis of the feature distribution's singular values confirms this profound anisotropy, a classic indicator of representation degeneration (Gao et al., 2019). We posit that this anisotropy masks the intrinsic capability of the pre-trained features to discern data redundancy, as in such a space, geometric distance is no longer a faithful proxy for semantic similarity. We argue that the failure of prior methods to account for this anisotropy is a primary source of their excessive data selection overhead. Therefore, to effectively measure data redundancy, it is essential to first re-center the feature distribution to mitigate this anisotropic bias and reveal the underlying semantic structure.

Building on this insight, we introduce **PRISM**, a novel, training-free framework that redefines multimodal data selection. PRISM bypasses the complexities of existing proxy-based and training-based methods by establishing a third paradigm we term **Intrinsic Selection**. The philosophy of this paradigm is to unlock the latent potential of an MLLM's pre-trained representations by correcting, rather than circumventing, the flawed geometry of its feature space. At its core, PRISM counteracts representation anisotropy by performing an *implicit re-centering* of the visual feature distribution. It effectively nullifies the global, non-semantic shift in the raw embeddings. This re-centering corrects a well-behaved isotropic geometry to the feature space. Crucially, it is this restoration that allows the **intrinsic features** themselves to become a direct and reliable signal for identifying visual se-

mantic redundancy. With the geometric distortions removed, there is no need for external models or expensive gradient computations, enabling highly effective visual instruction selection with negligible computational cost.

We validate PRISM through extensive experiments on a diverse set of multimodal benchmarks, evaluating its efficacy against state-of-the-art selection methods. Our results demonstrate that an MLLM fine-tuned on a small, PRISM-selected dataset (`PRISM-Instruct-250K`) outperforms models trained on the full dataset while reducing training time by 70%. On VisionFlan, PRISM selects only 57K samples from 186K examples and still exceeds full-data training. Furthermore, we conduct analyses on *Cross-Model Generalization* and *Knowledge Retention*, demonstrating that PRISM generalizes effectively across different MLLM architectures and better preserves the model's linguistic capabilities compared to full-dataset training.

Our key contributions are as follows:

- We identify *representation anisotropy* in the visual features of MLLMs as the root cause of inefficient data selection and diagnose its primary symptom, a *Global Semantic Drift*, which distorts similarity metrics.
- We propose **PRISM**, a training-free framework that establishes a new **Intrinsic Selection** paradigm. By performing an *implicit re-centering* of visual features, PRISM corrects an isotropic geometry, unlocking the MLLM's own features for redundancy detection.
- Our experiments show PRISM sets a new state-of-the-art, reducing training time by 70% while yielding performance superior to models fine-tuned on the complete dataset.

2 Visual Instruction Selection

While numerous methods exist for visual instruction selection, a rigorous framework for evaluating their practical utility is often overlooked. Here we establish two fundamental principles that any effective selection strategy must satisfy to be considered viable in real-world applications. These principles serve as the foundation for a new quantitative metric to holistically assess selection efficacy.

Principle 1: Performance Fidelity

A selected data subset must yield model performance equal to or greater than that achieved with the full dataset.

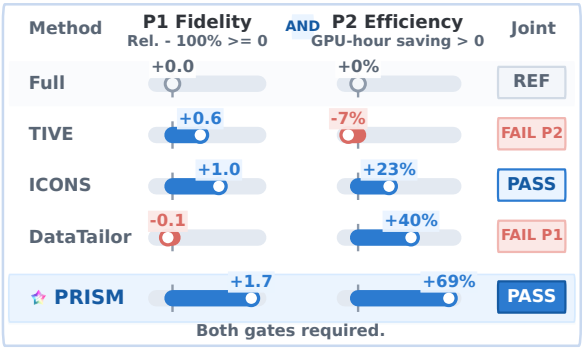

Figure 3: **Two-principle feasibility check.** A practical selector must satisfy both performance fidelity and net efficiency gain. PRISM is the only method that preserves full-data performance while substantially reducing end-to-end GPU hours.

Principle 2: Net Efficiency Gain

The total time consumed by data selection and subsequent model tuning must be strictly less than the time required for tuning on the full dataset.

To unify these principles into a single, quantifiable measure, we introduce the **Overall Selection Cost (OSC)** score. This metric captures the total overhead of a selection method in terms of both performance and computation. Formally, let $\mathcal{D}_{\text{full}}$ be the full instruction tuning dataset and $\mathcal{D}_{\text{sub}}$ be the selected subset. Let $P(\mathcal{D})$ denote the final performance score on the downstream benchmark achieved by a model fine-tuned on a dataset $\mathcal{D}$. The fine-tuning time on the full dataset is $T_{\text{tune}}(\mathcal{D}_{\text{full}})$, while the total time for a selection pipeline is the sum of the selection time T_{select} and the subset fine-tuning time $T_{\text{tune}}(\mathcal{D}_{\text{sub}})$. The Overall Selection Cost (OSC) score, $\mathcal{C}$, is defined as:

$$\mathcal{C} = \underbrace{\left(\frac{P(\mathcal{D}_{\text{full}})}{P(\mathcal{D}_{\text{sub}})}\right)}_{\text{Performance Ratio}} \times \underbrace{\left(\frac{T_{\text{select}} + T_{\text{tune}}(\mathcal{D}_{\text{sub}})}{T_{\text{tune}}(\mathcal{D}_{\text{full}})}\right)}_{\text{Time Ratio}}. \tag{1}$$

A selection method is considered practically viable if and only if its OSC score $\mathcal{C} < 1$. This single criterion guarantees that any computational savings are not achieved at the expense of final model performance, thus certifying a true net gain in efficiency as shown in Fig. 3 and Fig. 11.

As illustrated in Fig. 4, existing visual instruction selection methods can be broadly categorized into two main paradigms. The first is **Proxy-Based Selection**, where the target MLLM remains a black box, and data quality is evaluated using an external

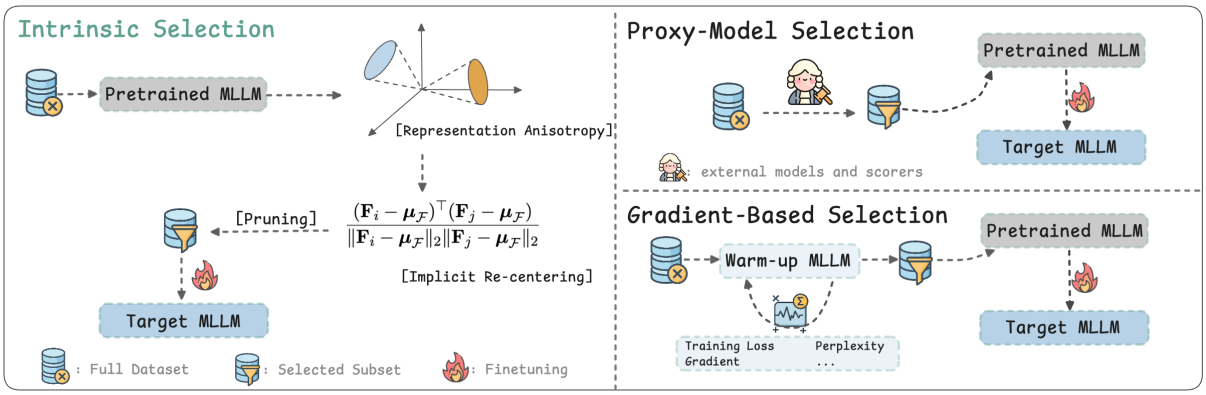

Figure 4: **A Comparison of Data Selection Paradigms.** Existing methods, such as Proxy-Based and Training-Based Selection, introduce significant computational overhead by relying on external scorers or expensive, iterative training loops to estimate data importance. In contrast, our proposed **Intrinsic Selection** paradigm operates on a more fundamental principle: we first diagnose a geometric flaw—*representation anisotropy*—in the MLLM's native visual features. We then apply an *implicit re-centering* to correct this bias. This isotropic restoration unlocks the intrinsic semantic structure of the data, enabling highly effective and training-free pruning in a single pass.

scoring function. This function can be a purpose-built scoring model (Chen et al., 2024a), a human reward model, or a powerful LLM-based scorer (Wei et al., 2023). Some approaches also train a smaller proxy MLLM to guide the data selection for a larger target model (Lee et al., 2024). The second category is **Training-Based Selection**. In this approach, the target MLLM is actively used to assess data value by computing metrics derived from its training dynamics, such as training loss, perplexity, or gradient information on a given data partition. Yet both paradigms exhibit structural drawbacks: they risk **performance degradation**, incur **high computational costs**, and introduce **Proxy-Based bias**. A fuller discussion appears in Appendix F. We contend that these drawbacks share a single, underappreciated root cause: the overlooked *representation anisotropy* inherent in MLLM visual feature distributions. To address this geometric pathology, we introduce **PRISM**—a self-**PR**uning **I**ntrinsic **S**election **M**ethod—that enables training-free multimodal data selection by operating directly on the model's intrinsic, re-centered features.

3 PRISM

3.1 Representation Anisotropy

To reveal the potential of pre-trained intrinsic features, we first investigate the feature distribution and uncover a critical, yet previously overlooked property in previous data selection methods: representation anisotropy. Representation anisotropy is a geometric pathology observed in many high-dimensional embedding spaces (Levi and Gilboa, 2024; Zhang et al., 2020). In an *isotropic* embedding, vectors are uniformly distributed around the origin and exhibit equal variance along every dimension; mathematically, the covariance matrix of an isotropic set is proportional to the identity. Such a space fully exploits its dimensionality and provides maximal directional diversity for encoding semantic concepts. As illustrated in Figure 2, we begin with a deep analysis of the feature geometry from a pre-trained MLLM, revealing a critical flaw in its visual representations. We find a stark disparity between modalities: while textual features are well-centered around a zero mean, the feature distributions for all visual datasets exhibit a significant, non-zero mean. This confirms that visual embeddings are confined to a narrow cone in high-dimensional space—a geometric bias that systematically distorts similarity metrics and gives rise to a phenomenon we term *Global Semantic Drift*, where shared background context is misinterpreted as semantic content. A singular value analysis of these distributions further confirms this profound anisotropy. The sharp "elbow point" observed in Figure 2 (right) indicates that feature variance is concentrated in only a few dominant dimensions, a classic symptom of representation degeneration that corrupts semantic distance measures (Levi and Gilboa, 2024). We posit that this unaddressed anisotropy is the primary reason for the inefficiency of prior art.

3.2 Theoretical Analysis: Anisotropy as a Source of Computational Overhead

To formalize this, let $\phi : \mathcal{D} \to \mathbb{R}^d$ be an embedding function mapping data points to representation vectors. The observed anisotropy implies a non-zero mean vector $\boldsymbol{\mu} = \mathbb{E}[\phi(d)]$ with significant magnitude. We can thus decompose any representation

$\mathbf{x}_i = \phi(d_i)$ into a shared global component and a sample-specific residual:

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\delta}_i \tag{2}$$

where $\boldsymbol{\delta}_i$ encodes the unique semantics of sample d_i. The defining characteristic of anisotropy is that the shared component dominates, i.e., $\|\boldsymbol{\mu}\|_2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2]$, which is empirically substantiated in Appendix H. This geometric property has a catastrophic consequence for standard metrics.

Theorem 1. *In an anisotropic representation space where $\|\boldsymbol{\mu}\|_2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2]$, the cosine similarity between any two randomly sampled vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is dominated by the global drift $\boldsymbol{\mu}$, masking their semantic dissimilarity encoded in $\boldsymbol{\delta}_i, \boldsymbol{\delta}_j$.*

Proof. The cosine similarity is given by:

$$\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{(\boldsymbol{\mu} + \boldsymbol{\delta}_i)^\top (\boldsymbol{\mu} + \boldsymbol{\delta}_j)}{\|\boldsymbol{\mu} + \boldsymbol{\delta}_i\|_2 \|\boldsymbol{\mu} + \boldsymbol{\delta}_j\|_2} \tag{3}$$

The numerator expands to $\|\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^\top(\boldsymbol{\delta}_i + \boldsymbol{\delta}_j) + \boldsymbol{\delta}_i^\top \boldsymbol{\delta}_j$. The denominator can be approximated using a first-order Taylor expansion $\sqrt{a+b} \approx \sqrt{a} + \frac{b}{2\sqrt{a}}$ for small b. As $\|\boldsymbol{\mu}\|_2$ is the dominant term, we have $\|\boldsymbol{\mu} + \boldsymbol{\delta}_i\|_2 \approx \|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_i}{\|\boldsymbol{\mu}\|_2}$. Substituting these into the main equation and retaining only the dominant terms yields:

$$\begin{aligned} \cos(\mathbf{x}_i, \mathbf{x}_j) &\approx \frac{\|\boldsymbol{\mu}\|_2^2}{\left(\|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_i}{\|\boldsymbol{\mu}\|_2}\right)\left(\|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_j}{\|\boldsymbol{\mu}\|_2}\right)} \\ &\approx 1 - \frac{1}{2}\left\| \frac{\boldsymbol{\delta}'_i}{\|\boldsymbol{\mu}\|_2} - \frac{\boldsymbol{\delta}'_j}{\|\boldsymbol{\mu}\|_2} \right\|_2^2 \approx 1 \end{aligned} \tag{4}$$

where $\boldsymbol{\delta}'$ is the component of $\boldsymbol{\delta}$ orthogonal to $\boldsymbol{\mu}$. The true semantic similarity, proportional to $\boldsymbol{\delta}_i^\top \boldsymbol{\delta}_j$, is relegated to lower-order terms, rendering the metric ineffective for distinguishing unique samples. A full derivation is provided in Appendix G. □

Corollary 2. *Due to Theorem 1, computationally efficient data selection algorithms predicated on geometric dissimilarity in the raw embedding space (e.g., Farthest-Point Sampling, distance-based clustering) fail to select semantically diverse subsets.*

This failure of inexpensive geometric proxies forces existing methods to adopt computationally intensive strategies to approximate the true, underlying data value, effectively paying a high computational "tax" for ignoring the geometric flaw of the space.

3.3 Redundancy Identification via Implicit Re-centering of Intrinsic Features

To solve the Representation Anisotropy and alleviate Global Semantic Drift, PRISM employs a implicit re-centering operation to restore a well-behaved isotropic geometry to the feature space. Crucially, it is this restoration that allows the intrinsic features themselves to become a direct and reliable signal for identifying semantic redundancy.

Let our dataset be $\mathcal{D} = \{d_1, \ldots, d_N\}$. For each data point d_i, we first extract a high-level feature representation $\mathbf{F}_i \in \mathbb{R}^d$ from an intermediate layer of pretrained MLLM. The set of these vectors, $\{\mathbf{F}_i\}_{i=1}^N$, forms the anisotropic space diagnosed in Section 3.2, characterized by the non-zero global mean vector $\boldsymbol{\mu}_\mathcal{F} = \frac{1}{N}\sum_{i=1}^N \mathbf{F}_i$. To measure similarity in a geometrically corrected space where the influence of this global drift is nullified, we keep the normalization-and-inner-product form of a Pearson-style correlation coefficient while adapting its re-centering term to the dataset-level mean. Specifically, we define the centered vector as $\mathbf{F}'_i = \mathbf{F}_i - \boldsymbol{\mu}_\mathcal{F}$, where $\boldsymbol{\mu}_\mathcal{F}$ captures the shared global shift across the corpus.

$$\rho(\mathbf{F}_i, \mathbf{F}_j) = \frac{(\mathbf{F}_i - \boldsymbol{\mu}_\mathcal{F})^\top (\mathbf{F}_j - \boldsymbol{\mu}_\mathcal{F})}{\|\mathbf{F}_i - \boldsymbol{\mu}_\mathcal{F}\|_2 \|\mathbf{F}_j - \boldsymbol{\mu}_\mathcal{F}\|_2} \tag{5}$$

This operation measures the alignment of two vectors' internal variation patterns after removing the global, non-semantic shift that characterizes anisotropy. By operating on these centered vectors $\mathbf{F}'_i$, we effectively evaluate the relationship between the unique semantic components, $\boldsymbol{\delta}_i$, that our theory identified.

With a robust similarity metric established, we quantify the redundancy of each sample. A sample is considered redundant if its unique semantic information is highly correlated with many other samples in the dataset. We define the *Redundancy Score* $\mathcal{R}(d_i)$ for a sample d_i as its expected correlation with the rest of the corpus:

$$\mathcal{R}(d_i) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \rho(\mathbf{F}_i, \mathbf{F}_j) \tag{6}$$

$\mathcal{R}(d_i)$ is a corpus-wide redundancy estimator: after re-centering, it measures the weighted degree of d_i in the complete semantic-similarity graph induced by $\rho(\cdot,\cdot)$. The score is computed by an exact aggregate implementation in $\mathcal{O}(Nd)$ time without materializing any $N \times N$ matrix; details in Appendix M. Given a selection budget defined by a percentile threshold τ, we select all samples whose redundancy scores fall below the τ-th percentile of the score distribution:

$$\mathcal{D}_{\text{select}} = \left\{ d_i \in \mathcal{D} \mid \mathcal{R}(d_i) \leq Q_\tau\left(\{\mathcal{R}(d_k)\}_{k=1}^N\right) \right\}, \tag{7}$$

Algorithm 1 PRISM

Input: Dataset $\mathcal{D} = \{d_1, \ldots, d_N\}$, MLLM components (VE, Proj, $\text{LLM}^{(l)}$), percentile τ
Output: Selected subset $\mathcal{D}_{\text{select}}$
Anisotropic Feature Extraction:
for all $d_i = (I_i, T_i) \in \mathcal{D}$ **do**
  $v_i \leftarrow \text{VE}(I_i)$
  $z_i \leftarrow \text{Proj}(v_i)$
  $\mathbf{F}_i \leftarrow \frac{1}{T}\sum_{t=1}^{T} \text{LLM}^{(l)}(z_i)_t$ /* Obtain layer-l representation */
end for
Redundancy Identification via Implicit Re-centering:
$\boldsymbol{\mu}_{\mathcal{F}} \leftarrow \frac{1}{N}\sum_{k=1}^{N} \mathbf{F}_k$
for all $i \in \{1, \ldots, N\}$ **do**
  $\mathcal{R}_i \leftarrow \frac{1}{N-1}\sum_{j \neq i} \frac{(\mathbf{F}_i - \boldsymbol{\mu}_{\mathcal{F}})^\top (\mathbf{F}_j - \boldsymbol{\mu}_{\mathcal{F}})}{\|\mathbf{F}_i - \boldsymbol{\mu}_{\mathcal{F}}\|_2 \|\mathbf{F}_j - \boldsymbol{\mu}_{\mathcal{F}}\|_2}$ /* Average correlation is implemented by an exact aggregate form in $\mathcal{O}(Nd)$ */
end for
Diversity-Aware Pruning:
$q_\tau \leftarrow \text{Percentile}(\{\mathcal{R}_k\}_{k=1}^{N}, \tau)$ /* Find score threshold at τ-th percentile */
$\mathcal{D}_{\text{select}} \leftarrow \{d_i \in \mathcal{D} \mid \mathcal{R}_i \leq q_\tau\}$ /* Filter samples with low redundancy scores */
return $\mathcal{D}_{\text{select}}$

where $Q_\tau(\cdot)$ is the function that computes the τ-th percentile. This principled approach ensures that the selected subset is enriched with semantically diverse and informative samples, having explicitly corrected for the geometric distortions of the original embedding space.

4 Experiments

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. To test robustness beyond the LLaVA-style mixture, we additionally evaluate PRISM on VisionFlan-186K (Xu et al., 2024), a distinct instruction pool containing 191 vision-language tasks and 186K samples. 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 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., 2024a), 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 rigorous framework of LLaVA-1.5 (Liu et al., 2024a), we evaluate PRISM across a comprehensive suite of seven diverse benchmarks. This suite is designed to probe a wide array of MLLM capabilities, including fine-grained perception, visual reasoning, hallucination detection, and complex conversational skills. Our evaluation includes MMBench (Liu et al., 2024c), ScienceQA (Lu et al., 2022), MME (Fu et al., 2025), POPE (Li et al., 2023), VizWiz (Gurari et al., 2018), MM-Vet (Yu et al., 2024b), and MMMU (Yue et al., 2023). Detailed descriptions of each benchmark are provided in Appendix I.

4.2 Main Results

We present a comprehensive evaluation to validate PRISM's effectiveness and robustness. Our primary results in Table 1 demonstrate that PRISM outperforms both full fine-tuning and state-of-the-art selection methods on the LLaVA-1.5-7B model. We then showcase PRISM's broad applicability by demonstrating its strong performance across different MLLM architectures (Table 8), in uni-modal, text-only settings (Fig. 6 and Table 5), and on the distinct VisionFlan instruction pool (Fig. 6 and Appendix B.4). We highlight the following findings:

① **State-of-the-Art Performance with Less Data.** As shown in Table 1, PRISM achieves superior performance across various benchmarks, outperforming full-dataset fine-tuning by a relative **1.7%**. The gains are particularly notable on complex reasoning tasks like MMBench (65.2 vs. 64.3) and MM-Vet (32.0 vs. 31.1). This demonstrates that PRISM preferentially selects high-value, instruction-rich data more effectively than existing methods.

② **Cross-Model Generalization and Scalability.** PRISM's selection mechanism generalizes robustly across diverse MLLM architectures by correcting representation anisotropy. By identifying semantically dense samples, we establish the approach's breadth across five model–encoder combinations. Results demonstrate that effectiveness holds for both small and large language models—including Vicuna, Phi-2, Qwen2.5, and Llama-3—and for dis-

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 better 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.

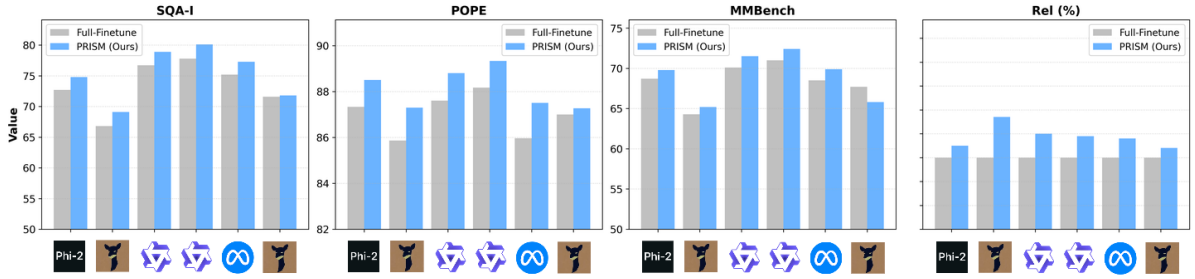

Figure 5: **Generalization across diverse MLLM families and scales.** Comparison of PRISM (blue) and full fine-tuning (grey) across SQA-I, POPE, MMBench, and Rel. (%) (detailed numerical results provided in Table 8). Results for **Phi-2**, **Vicuna**, **Llama-3**, and **Qwen2.5** confirm that PRISM's performance gains are architecture-agnostic and transfer robustly across varied backbones and vision encoders.

tinct vision encoders like CLIP and SigLIP (details in Appendix 7). These findings, illustrated in Fig.5, confirm that PRISM's gains are not specific to the LLaVA architecture but transfer effectively across diverse multimodal families.

③ **Balanced Efficiency and Performance.** As illustrated in Figure 8, PRISM occupies the most favorable efficiency–performance region: it improves over full fine-tuning while substantially reducing end-to-end GPU hours. This advantage comes from its training-free design and aggregate correlation computation. A detailed breakdown of the wall-clock runtime can be found in Appendix M.

④ **Robustness to Instruction-Pool Shift.** Beyond LLaVA-665K, PRISM also generalizes to VisionFlan-186K, a more task-diverse instruction pool with 191 distinct vision-language tasks. With the same 30% selection budget, PRISM selects 57K samples and reaches 100.9% relative performance over full-data training, substantially outperforming random selection at the same budget (94.1%). The comparison is summarized in Fig. 6, with full results in Table 6. This confirms that PRISM's geometric selection principle is not tied to a single LLaVA-style data mixture.

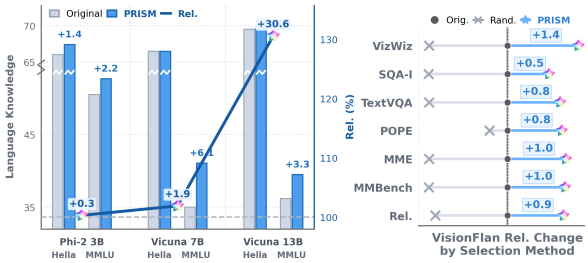

Figure 6: PRISM preserves language knowledge and robustly generalizes to VisionFlan-186K Instruction Pool.

4.3 In-Depth Model Behavior Analysis

♥ **Effective Hallucination Mitigation.** PRISM sets a new state-of-the-art on all POPE subsets for hallucination detection, surpassing even specialized methods. This result suggests that our core mechanism—alleviating representation anisotropy via implicit re-centering —is highly effective at removing samples containing spurious semantic drift, which are a known cause of object hallucination.

♠ **Language Knowledge Retention.** Visual instruction tuning is known to degrade a model's foundational language capabilities, a phenomenon often termed "knowledge forgetting" (Zhang et al., 2024). We show that PRISM not only mitigates this effect but can even enhance text-only perfor-

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 2: **Ablation study validating the core design principles of PRISM.** The results confirm that performance is maximized by using Shallow Layer features (preserving a cleaner geometric structure), selecting for Low Correlation (maximizing informational diversity), and applying Average Pooling (ensuring a holistic representation).

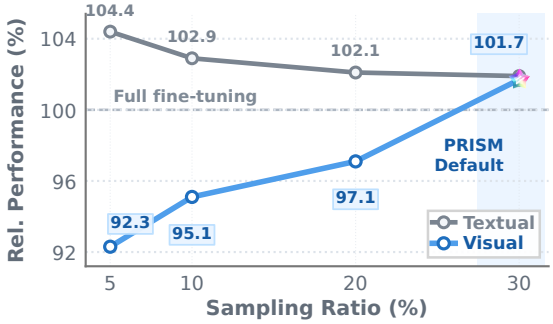

Figure 7: **Sampling-ratio trade-off in PRISM.** Visual performance (blue) increases with larger selected subsets and surpasses full fine-tuning at 30%, while textual performance is best preserved under smaller budgets.

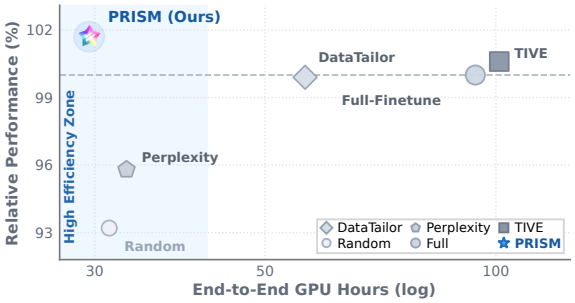

Figure 8: **Efficiency–performance Pareto comparison.** PRISM achieves the highest relative performance (101.7%) while reducing total GPU time by about 70% compared with full fine-tuning. TIVE attains comparable accuracy but exceeds the full fine-tuning cost, whereas cheaper heuristics such as Random and Perplexity sacrifice substantial performance.

mance. We hypothesize this is because PRISM's Pearson-style correlation scoring effectively removes samples with spurious visual-textual associations. By filtering out this "noisy" data that often causes catastrophic forgetting, our method provides a cleaner training signal. As shown in Fig. 6 and detailed in Table 5, models trained on the PRISM subset consistently outperform their full-dataset-trained counterparts on language-centric benchmarks. This highlights a dual benefit of PRISM: achieving state-of-the-art multimodal performance without sacrificing language competency.

♦ **Analysis of the Sampling Ratio.** We analyze PRISM under 4 sampling ratios to characterize the trade-off between data budget and dual-modality performance (Figure 7). The results show that PRISM remains robust across budgets, with larger subsets favoring visual performance and smaller subsets better preserving language knowledge.

4.4 Ablation Study

To validate the design choices of PRISM, we conduct ablation studies on three key components: the layer of visual features, the Pearson-style correlation scoring, and the aggregation strategy.

① **Influence of LLM Layer Selection.** We find that features from a **Shallow Layer** (Layer 1) are most effective for PRISM, significantly outperforming those from middle or deep layers (Table 2). This suggests that the fundamental geometric properties required for redundancy detection are most distinct in early layers, before complex abstractions introduce further representational artifacts.

② **Impact of Correlation-based Selection.** Validating our central hypothesis, we show that selecting the **Low-Correlation** data subset yields the highest performance. This confirms that our implicit re-centering strategy effectively identifies informational diversity and prunes redundant samples that provide diminishing returns for model training (see Appendix F for further analysis).

③ **Effect of Token Aggregation Strategy.** We confirm that using an **Average** of all image tokens to form the global feature vector is superior to using only the **Last Image Token**. Averaging provides a more holistic and robust representation of an image's semantics, which is crucial for the stability of our geometric analysis, whereas relying on a single token can be susceptible to positional biases.

5 Conclusion

We trace MLLM selection inefficiency to representation anisotropy and propose **PRISM**, a training-free method that corrects semantic structure via implicit re-centering. **PRISM** reduces training time by 70% while surpassing full-dataset performance.

Limitations

While PRISM demonstrates significant promise, we acknowledge several limitations that offer avenues for future research. Our current approach focuses specifically on identifying and pruning semantic redundancy based on feature correlation. It is not designed to detect other critical data issues, such as factual inaccuracies, ethical biases within the instruction data. Future work could explore extending our geometry-aware principles to address these other facets of data quality or adapting the method to new modalities beyond vision and language.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025a. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15230–15250, Vienna, Austria. Association for Computational Linguistics.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hin rich Schuetze, Volker Tresp, and Yunpu Ma. 2025b. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. *ArXiv*, abs/2505.13408.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. 2023. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *Preprint*, arXiv:2305.09246.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024a. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *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, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *ArXiv*, abs/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*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *Preprint*, arXiv:1907.12009.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xingyue Huang, Rishabh, Gregor Franke, Ziyi Yang, Jiamu Bai, Weijie Bai, Jinhe Bi, Zifeng Ding, Yiqun Duan, Chengyu Fan, Wendong Fan, Xin Gao, Ruohao Guo, Yuan He, Zhuangzhuang He, Xianglong Hu, Neil Johnson, Bowen Li, Fangru Lin, and 27 others. 2025. Loong: Synthesize long chain-of-thoughts at scale through verifiers. *Preprint*, arXiv:2509.03059.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2021. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

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

Meir Yossef Levi and Guy Gilboa. 2024. The double-ellipsoid geometry of clip. *arXiv preprint arXiv:2411.14517*.

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. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Ziming Li, Xiaoming Wu, Zehong Wang, Jiazheng Li, Yijun Tian, Jinhe Bi, Yunpu Ma, Yanfang Ye, and Chuxu Zhang. 2026. Graph is a substrate across data modalities. *Preprint*, arXiv:2601.22384.

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 *Advances in Neural Information Processing Systems*.

Rui Lu, Jinhe Bi, Yunpu Ma, Feng Xiao, Yuntao Du, and Yijun Tian. 2025. Mv-debate: Multi-view agent debate with dynamic reflection gating for multimodal harmful content detection in social media. *Preprint*, arXiv:2508.05557.

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.

Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, Feng Xiao, and Lizhen Cui. 2025. Can visual input be compressed? a visual token compression benchmark for large multimodal models. *Preprint*, arXiv:2511.02650.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763.

Yijun Tian, Shaoyu Chen, Zhichao Xu, Yawei Wang, Jinhe Bi, Peng Han, and Wei Wang. 2025. Reinforcement mid-training. *Preprint*, arXiv:2509.24375.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *ArXiv*, abs/2406.16860.

Guancheng Wan, Lucheng Fu, Haoxin Liu, Yiqiao Jin, Hui Yi Leong, Eric Hanchen Jiang, Hejia Geng, Jinhe Bi, Yunpu Ma, Xiangru Tang, B. Aditya Prakash, Yizhou Sun, and Wei Wang. 2025a. Beyond magic words: Sharpness-aware prompt evolving for robust large language models with tare. *Preprint*, arXiv:2509.24130.

Guancheng Wan, Xiaoran Shang, Yuxin Wu, Guibin Zhang, Jinhe Bi, Liangtao Zheng, Xin Lin, Yue Liu, Yanbiao Ma, Wenke Huang, and Bo Du. 2025b. HYPERION: Fine-grained hypersphere alignment for robust federated graph learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15271–15342, Bangkok, Thailand. Association for Computational Linguistics.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Fang Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, and 8 others. 2024. Blip-3: A family of open large multimodal models. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 6183–6194.

Minglai Yang, Xinyu Guo, Zhengliang Shi, Jinhe Bi, Steven Bethard, Mihai Surdeanu, and Liangming Pan. 2026. Alignsae: Concept-aligned sparse autoencoders. *Preprint*, arXiv:2512.02004.

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, and 3 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206.

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 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., 2023; Chen et al., 2023; Xia et al., 2024; Tian et al., 2025; Lu et al., 2025; Bi et al., 2025b; Peng et al., 2025; Wan et al., 2025b,a; Huang et al., 2025; Li et al., 2026; Yang et al., 2026). State-of-the-art models like BLIP3 (Xue et al., 2024), InternVL2.5 (Chen et al., 2024b), 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., 2024a) 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.

B Dataset and Selection Details

This section provides a detailed breakdown of our `PRISM-Instruct-250K` dataset and the underlying data distributions that informed its creation. The dataset is the result of applying the PRISM selection methodology to the LLaVA-665K data pool, which aggregates samples from diverse sources. Our goal is to create a compact, high-utility subset by retaining the most informative, low-redundancy samples.

B.1 Redundancy Score Distribution

At the core of our selection process is the *Redundancy Score* $\mathcal{R}(d_i)$, which quantifies the degree to which a sample is semantically similar to the rest of the corpus. Figure 10 visualizes the distribution of these scores (lower is better) for each source dataset within the LLaVA-665K pool.

The distributions reveal the varying levels of internal diversity within each source. For instance, datasets like LLaVA and VG exhibit a wider distribution with a significant number of low-score (high-value) samples. In contrast, other datasets may have a higher density of redundant, high-score samples. PRISM applies a single, global threshold to this combined pool, ensuring that only the top 30% of samples—those with the lowest redundancy scores, regardless of their origin—are selected. This principled approach prevents any single, potentially redundant dataset from dominating the final selection and instead curates a subset based on universal informational value.

Scope of re-centering. We emphasize that PRISM does not claim to perform full whitening or to make the empirical covariance matrix exactly proportional to the identity. Instead, PRISM targets the dominant first-order component of visual representation anisotropy: the corpus-level mean shift. Our analysis shows that this global component can dominate raw cosine similarity and obscure sample-specific semantic variation. Therefore, the goal of PRISM is not to completely isotropize the representation space, but to remove the shared global drift that is most harmful for redundancy estimation.

Why mean re-centering is sufficient for selection. For the purpose of training-free data selection, full covariance correction is not always necessary. Redundancy ranking only requires a similarity signal that better reflects relative sample-specific variation. By subtracting the corpus mean and comput-

Top-Ranked Sample

[Question] What region of the world is this located?
[Answer] Asia
[Reasoning] Requires fine-grained OCR and cultural reasoning.

Bottom-Ranked Sample

[Question] What type of weather is this?
[Answer] Sunny
[Reasoning] Trivial perception; high semantic redundancy.

Figure 9: **Representative examples selected by PRISM from LLaVA-665K.** Top-ranked samples correspond to more challenging and informative interactions that exhibit unique semantic structures, whereas bottom-ranked samples tend to be trivial, binary, or semantically redundant.

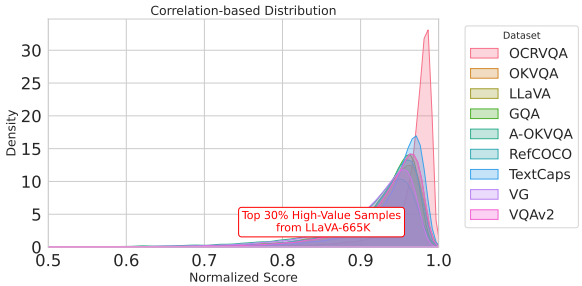

Figure 10: **Distribution of PRISM's Redundancy Scores across source datasets.** The vertical line indicates the global selection threshold used to prune the 70% most redundant samples (i.e., those with the highest correlation scores). Datasets with a higher density of low-correlation samples contribute more instances to the final `PRISM-Instruct-250K` set, ensuring a merit-based composition.

ing normalized inner products in the centered space, PRISM suppresses the common background direction while preserving an efficient one-pass scoring pipeline. This design deliberately trades complete second-order correction for scalability, numerical stability, and ease of deployment on large multimodal instruction pools.

Why not full whitening or principal-component removal? A natural alternative is to explicitly whiten the representation space or remove the top principal components. We view these methods as complementary but less aligned with our objective. First, whitening requires reliable covariance estimation in a high-dimensional, low-effective-rank regime, where the covariance matrix can be ill-conditioned and requires additional regularization choices. Second, top-PC removal introduces an extra hyperparameter: the number of components to remove. This hyperparameter can be dataset- and model-dependent, reducing the plug-and-play nature of training-free selection. Third, both approaches increase implementation complexity compared with the corpus-mean re-centering used by PRISM. PRISM therefore focuses on the first-order global drift, which yields a strong efficiency–performance trade-off without additional tuning.

B.2 Final Dataset Composition

The application of our selection strategy results in the `PRISM-Instruct-250K` dataset. Table 4 details the final number of samples drawn from each source. The composition reflects the data-driven nature of our method: sources rich in unique, low-correlation samples are more heavily represented. The result is a balanced and diverse instruction-tuning dataset that has been algorithmically optimized for both performance and efficiency.

B.3 Language Knowledge Retention Details

Table 5 reports the text-only evaluation underlying Fig. 6. Across three model scales, PRISM preserves or improves language knowledge after visual instruction tuning, with the largest gain observed on the 13B Vicuna backbone.

Method	Feature Space	Corrects Mean Drift?	Extra Hyperparam.?	Scalable?	Role / Limitation
Random selection	None	No	No	Yes	Matched-budget lower-bound control. It tests whether gains come merely from using fewer samples, but it does not exploit visual geometry or redundancy.
Raw cosine redundancy	Target MLLM features	No	No	Yes	Direct ablation of PRISM's re-centering. It remains dominated by the corpus-level drift in anisotropic visual spaces and therefore measures shared background directions rather than sample-specific semantic variation.
PRISM	Target MLLM features	Yes	No	Yes	Our default intrinsic selector. It removes the first-order corpus-level shift and computes redundancy through an exact aggregate form without constructing an $N \times N$ similarity matrix.
Top-PC removal + cosine	Target MLLM features	Partially	Yes: number of PCs k	Moderate	A reasonable diagnostic alternative, but it introduces a dataset/model-dependent choice of k. Removing too few PCs may leave drift; removing too many may discard semantic directions useful for instruction tuning.
Whitening / Mahalanobis-style correction	Target MLLM features	Yes, plus covariance correction	Yes: shrinkage / rank / clipping	No / costly	More aggressive second-order correction, but requires estimating and inverting a high-dimensional covariance matrix. In low-effective-rank visual spaces, this is ill-conditioned and requires regularization, weakening plug-and-play training-free deployment.
Centered k-center / farthest-point sampling	Target MLLM features	Yes, if centered	Yes: distance metric / approximation	No for exact greedy	Optimizes geometric coverage rather than corpus-wide redundancy degree. Exact greedy selection is expensive at LLaVA-665K scale; approximate variants introduce additional engineering and approximation choices.
CLIP-feature selection	External / vision-encoder feature space	Depends on preprocessing	Possibly	Yes	Tests whether generic vision features are sufficient. However, it is not an intrinsic selector over the target MLLM's post-projection representation, so it may be misaligned with the actual instruction-tuning model.

Table 3: **Applicability analysis of simple geometry baselines.** We compare PRISM with natural geometry-based alternatives under the desiderata of training-free multimodal instruction selection: (i) correcting the dominant global drift, (ii) avoiding additional tuning, (iii) preserving linear or near-linear scalability, and (iv) operating on the target MLLM's intrinsic feature space. These alternatives are useful diagnostic controls, but most fail at least one desideratum.

Data Source	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 4: Final sample distribution of the `PRISM-Instruct-250K` dataset after applying our selection methodology to the LLaVA-665K pool.

Model	HellaSwag	MMLU	MMLU-PRO	Rel. (%)
LLaVA-Phi2-3B	66.0	50.5	9.1	100.0
PRISM-3B	**67.4**	**52.7**	8.6	**100.3**
LLaVA-Vicuna-7B	66.5	35.0	17.8	100.0
PRISM-7B	**66.5**	**41.1**	15.7	**101.9**
LLaVA-Vicuna-13B	69.5	36.2	6.8	100.0
PRISM-13B	**69.6**	**39.5**	**12.4**	**130.6**

Table 5: Text-only performance after visual instruction tuning. Relative scores are normalized to each full-data model.

B.4 Robustness on VisionFlan

We further evaluate PRISM on VisionFlan-186K (Xu et al., 2024), a task-diverse instruction pool containing 191 distinct vision-language tasks and 186K samples. Unlike the LLaVA-665K mixture, VisionFlan preserves fine-grained task diversity and therefore provides a complementary test of whether PRISM remains effective under a different instruction distribution.

Using the same 30% selection budget, PRISM selects 57K samples from VisionFlan-186K. As shown in Table 6, full-data training reaches 100.0% relative performance, random selection with 57K samples reaches 94.1%, and PRISM reaches 100.9%. PRISM slightly exceeds the full-data baseline while using only about one third of the instruction pool.

C Selection Budget and Relative Score Calculation

Clarification of the 30% selection budget. The 30% budget used in our main experiments refers to the selection budget over image-associated multimodal samples, not over the entire LLaVA-665K mixture. Following the data composition protocol of LLaVA-1.5, the instruction-tuning mixture contains both visual instruction samples and text-only language samples. Since PRISM is designed to estimate redundancy from visual representations, we apply PRISM only to the visual subset and retain the lowest-redundancy 30% visual samples. The text-only samples are kept unchanged across

Method	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench	Rel. (%)
Full Data (186K)	41.7	60.8	50.4	83.4	1263.2	52.6	100.0
Random (57K)	38.8	56.5	46.9	83.1	1175.0	48.9	94.1
PRISM (57K)	**42.3**	**61.1**	**50.8**	**84.1**	**1275.5**	**53.1**	**100.9**

Table 6: Robustness evaluation on VisionFlan-186K. PRISM selects 57K samples and matches or exceeds full-data training across the aggregate score.

PRISM and all comparable baselines.

This design choice has two purposes. First, it isolates the effect of visual-instruction pruning, ensuring that performance differences are not caused by different amounts of language-only supervision. Second, it follows the original LLaVA-1.5 mixture structure, where language-only samples are included to preserve the language capability of the base model. As a result, the final PRISM-Instruct subset contains the selected 30% visual samples together with the unchanged text-only samples. In our LLaVA-665K setting, this yields 250,557 total training examples, corresponding to approximately 38% of the full mixture.

Fairness protocol for baseline comparison. For a fair comparison, all selection methods should follow the same data-budget rule: the selection budget is applied to the visual subset only, while the text-only subset is retained unchanged. In addition, all methods should use the same base model, visual instruction pool, training recipe, number of epochs, evaluation scripts, and, when applicable, random seed. For baselines reproduced in our pipeline, we enforce this protocol directly. For baselines adopted from prior work, we mark their provenance explicitly and only compare them when their reported setting matches the LLaVA-1.5-7B/LLaVA-665K setup as closely as possible.

Relative score calculation under missing entries. To avoid imputing unreported results, we compute the relative score only over the set of available metrics for each method. Specifically, let $\mathcal{M}_s$ denote the set of metrics reported for method s, and let $P_{s,m}$ and $P_{\text{full},m}$ denote the score of method s and the full-data fine-tuning baseline on metric m, respectively. Since all metrics in our table are higher-is-better, the relative score is computed as

$$\text{Rel.}(s) = 100 \times \frac{1}{|\mathcal{M}_s|} \sum_{m \in \mathcal{M}_s} \frac{P_{s,m}}{P_{\text{full},m}}.$$

For metrics with multiple subsets, such as POPE-P, POPE-R, and POPE-A, each reported subset is treated as one metric. If a prior work only reports the average over POPE subsets, we use the reported average only for that baseline and mark this in the table caption.

Because methods with missing entries are averaged over different metric sets, their relative scores should be interpreted as available-metric relative performance rather than a fully matched aggregate across all benchmarks. Our main claims about PRISM are based on the complete metric set, since PRISM and the full-data baseline are evaluated on all benchmarks under the same protocol.

D Model Architecture Details and Generalization Setup

To rigorously evaluate the generalization capabilities of PRISM, we extend our analysis across a diverse set of Multimodal Large Language Models (MLLMs). The objective is to verify whether the visual anisotropy diagnosis and the subsequent performance gains from PRISM remain consistent across different model scales, language model families, and vision encoder architectures. This broad evaluation ensures that the intrinsic semantic properties identified by PRISM are not artifacts of a specific model configuration but represent fundamental characteristics of multimodal data representations.

As summarized in Table 7, our evaluation framework encompasses the following dimensions:

- **Language Model Families and Scales**: Beyond the Vicuna and Phi-2 families discussed in the main text, we include the Qwen2.5 series (7B Base and Instruct) and Llama-3 (8B) to represent state-of-the-art open-source LLM backbones.

- **Vision Encoders**: We evaluate models utilizing various vision foundation backbones, specifically CLIP ViT-L/14 (336px) and SigLIP-SO400M, to ensure PRISM's robustness to different cross-modal alignment strategies.

- **Generalization Analysis**: For each configuration, we conduct both a systematic anisotropy diagnosis (as detailed in Section 3.2) and an end-to-end evaluation of the performance improvements yielded by PRISM-selected data.

Model Family	Vision Encoder	Language Model Backbone
LLaVA-Vicuna-7B	CLIP ViT-L/14 (336px)	Vicuna v1.5 7B
LLaVA-Phi2-3B	SigLIP-SO400M	Phi-2 2.7B
LLaVA-Vicuna-13B	CLIP ViT-L/14 (336px)	Vicuna v1.5 13B
LLaVA-Qwen-7B	SigLIP-SO400M	Qwen2.5-7B (Base/Instruct)
LLaVA-Llama-8B	CLIP ViT-L/14 (336px)	Llama-3-8B

Table 7: **Comprehensive architectural configurations for generalization testing.** The evaluation spans multiple LLM families, varying parameter scales, and diverse vision encoders to validate the universal applicability of PRISM.

E Impact of Textual Representation and Joint Multimodal Selection

To address potential concerns regarding the role of textual prompts in data redundancy, this section provides: (1) a comparative analysis of visual vs. textual anisotropy, (2) the motivation for our visual-centric approach, and (3) additional experimental results explicitly incorporating textual features during selection.

E.1 Contrastive Anisotropy: Visual vs. Textual Modalities

Our diagnosis of representation anisotropy reveals a significant disparity between modalities. As illustrated in Figure 2, while visual features across all instruction datasets exhibit a non-zero mean vector μ with substantial magnitude, text features produced by the same MLLM are comparatively well-centered around the origin. Furthermore, the singular value spectrum of the textual embedding space is considerably flatter than that of the visual space.

Because these two modalities exhibit fundamentally different geometric behaviors, applying a single global re-centering operation to concatenated multimodal features would be suboptimal. A joint mean vector would be dominated by the severe visual drift, potentially over-correcting or distorting the near-isotropic textual geometry, thereby degrading the quality of similarity estimates. Consequently, PRISM focuses its geometric correction on the visual stream, where anisotropy is the primary obstacle to redundancy detection.

E.2 Primacy of Visual Redundancy in Instruction Tuning

In the context of visual instruction tuning, the image typically serves as the primary carrier of new scene-level semantics. In datasets such as LLaVA-665K, text prompts often function as task-specific templates (e.g., questions or instructions) conditioned on the visual content. Therefore, redundancy in these datasets is largely governed by the visual channel: models frequently encounter near-identical visual semantics under slight variations of textual phrasing. By targeting visual anisotropy, PRISM effectively prunes redundant visual scenes that provide diminishing returns during the fine-tuning process.

E.3 Additional Experiments with Multimodal Features

To empirically verify this design choice, we conducted additional experiments on LLaVA-1.5-7B using features that explicitly incorporate both visual and textual information. We constructed multimodal representations by aggregating image tokens and text prompt tokens before applying the PRISM selection pipeline.

As shown in Table 9, including textual prompts in the selection signal does not yield performance gains over our visual-only approach. In fact, the multimodal-feature selection (97.8% relative performance) underperforms the visual-only PRISM (101.7%). We hypothesize this is because text prompts introduce additional variance related to syntactic style and template length that is uncorrelated with the underlying semantic redundancy of the task. These results reinforce that visual anisotropy is the dominant factor to address for training-free intrinsic selection.

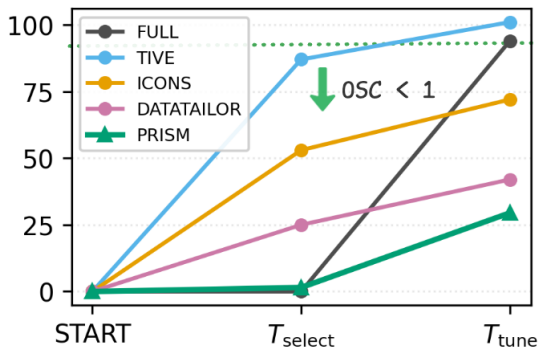

Figure 11: End-to-end computational cost (GPU Hours) vs. final model performance. PRISM achieves state-of-the-art performance while significantly reducing total time, thus satisfying the condition OSC < 1.

F Limitations of Existing Methods and Our Geometric Motivation

This section provides a detailed analysis of the structural limitations inherent in existing visual instruction selection paradigms and introduces the theoretical motivation for our work. We first outline the three primary failure modes of prior methods and then present our diagnosis that these issues stem from a single, unaddressed geometric flaw in MLLM feature representations.

F.1 A Critical Review of Existing Selection Paradigms

Despite their diverse approaches, current selection methods consistently exhibit three critical drawbacks that limit their practical utility, often violating the core principles of effective data selection we established in Section 2.

1. Performance Degradation. A primary goal of data selection is to curate a smaller, more potent dataset that ideally improves model performance. However, as shown in our main results (Table 1), many existing methods paradoxically lead to models that underperform a full fine-tuning baseline. This outcome directly contradicts our *Principle 1 (Performance Fidelity)*, suggesting that their heuristics for identifying "valuable" data are often misaligned with the true learning objectives.

2. Prohibitive Computational Cost. The second major failure mode is excessive computational overhead, particularly in Training-Based methods. For instance, techniques like TIVE (Liu et al., 2024d) require a costly "warm-up" phase, often involving LoRA-based (Hu et al., 2021; Bi et al., 2025a) fine-tuning, simply to generate the gradient signals needed for selection. As we show in our runtime analysis (Table 15), the time spent on selection alone can exceed the time saved during the subsequent tuning, violating *Principle 2 (Net Efficiency Gain)* and rendering the approach impractical at scale.

3. Proxy Model Bias. To avoid the high costs of gradient computation, some methods delegate scoring to an external proxy model (e.g., GPT-4 or a reward model). While faster, this introduces a significant bias, as the proxy's scoring function may not align with the target MLLM's specific learning needs. This approach yields a generic, one-size-fits-all dataset, failing to provide an optimal, tailored subset for the specific MLLM being trained.

F.2 The Unifying Cause: A Deep Dive into Representation Anisotropy

We argue that these seemingly disparate issues are symptoms of a single, deeper problem: a failure to account for the pathological geometry of MLLM feature spaces. In practice, high-dimensional embeddings learned by deep networks are rarely isotropic (uniformly distributed); instead, they are highly **anisotropic**, exhibiting a strong directional bias that corrupts the very notion of semantic distance and plagues models across modalities (Ethayarajh, 2019; Jing et al., 2021).

The Geometry of Anisotropy. An embedding space is said to be **isotropic** if its features are, on average, uniformly distributed in all directions from the origin. For a zero-mean feature set $\mathcal{X} = \{\mathbf{x}_i\}$, this property is formally defined by a covariance matrix proportional to the identity matrix, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \propto \mathbf{I}$. Such a space leverages its full dimensionality to represent concepts. In stark contrast, an **anisotropic** space is characterized by a non-uniform, biased distribution.

Low Effective Rank and Representation Degeneration. A complementary perspective comes from singular value decomposition (SVD). In an anisotropic space, the singular value spectrum of the centered feature matrix exhibits a sharp, steep decay. As illustrated in Figure 2, a few dominant singular values capture most of the data's variance, while the rest are negligible. This indicates a **low effective rank** (Arora et al., 2016), meaning the representation fails to utilize the full capacity of its embedding dimensions. This phenomenon, often called **representation degeneration**, is prevalent in modern architectures, including those trained with contrastive objectives like CLIP (Radford et al., 2021).

Implications for Data Selection. This unaddressed geometric flaw helps explain the practical limitations of prior art. Because simple and efficient geometric metrics (e.g., cosine similarity, farthest-point sampling) become unreliable under strong anisotropy, many existing methods turn to more expensive proxy signals to approximate semantic value. Training-Based methods pay a heavy price in computation to derive signals from the model's dynamics, while Proxy-Based methods introduce potential biases from an external scorer. Both are, in effect, paying a high computational "tax" for not first correcting the geometry of the

Model	SQA	SQA-I	VizWiz	POPE-P	POPE-R	POPE-A	MM-Vet	MMBench	MME	MMMU	Rel. (%)
LLaVA-Phi2-3B	75.3	72.7	41.2	87.3	88.6	86.1	35.6	68.7	1765.7	37.7	100%
PRISM-3B	**76.3**	**72.8**	40.9	**87.5**	**88.8**	**86.5**	34.1	**68.9**	**1790.5**	37.6	**100.1%**
LLaVA-Vicuna-7B	69.4	66.8	50.0	86.1	87.3	84.2	31.1	64.3	1822.6	35.4	100%
PRISM-7B	**71.3**	**69.1**	**50.1**	**87.7**	**88.7**	**85.5**	**32.0**	**65.2**	**1800.0**	34.7	**101.7%**
LLaVA-Vicuna-13B	74.4	71.6	53.6	87.4	88.0	85.6	36.1	67.7	1826.7	35.1	100%
PRISM-13B	**74.5**	**71.8**	53.1	**87.7**	**88.4**	**85.7**	**36.4**	65.8	**1846.0**	**35.7**	**100.4%**
LLaVA-Qwen-7B (Base)	79.1	76.7	55.3	87.3	89.1	86.4	33.1	70.1	1923.1	46.0	100%
PRISM-Qwen-7B (Base)	**81.2**	**78.9**	55.1	**88.6**	**90.3**	**87.5**	**33.7**	**71.5**	**1895.8**	45.2	**101.0%**
LLaVA-Qwen-7B (Instruct)	80.2	77.8	54.7	87.9	89.7	86.9	33.8	71.0	1935.4	46.8	100%
PRISM-Qwen-7B (Instruct)	**82.3**	**80.1**	54.2	**89.1**	**90.8**	**88.1**	**34.2**	**72.4**	**1908.6**	46.1	**100.9%**
LLaVA-Llama-8B	77.6	75.2	53.7	86.3	87.6	84.0	31.8	68.5	1889.7	40.3	100%
PRISM-Llama-8B	**79.4**	**77.3**	53.4	**87.8**	**89.1**	**85.6**	31.5	**69.9**	**1862.5**	39.7	**100.8%**

Table 8: **Cross-Model Generalization and Scalability of PRISM.** PRISM's selection mechanism generalizes robustly across diverse MLLM architectures by correcting representation anisotropy, as shown across five LLaVA-style model–encoder combinations. By identifying semantically dense samples for each architecture, PRISM consistently improves performance over full finetuning for both small and large language models—including Vicuna, Phi-2, Qwen2.5, and Llama-3—as well as distinct vision encoders such as CLIP and SigLIP. Together with Fig. 5, these results confirm that PRISM's gains are not tied to a particular LLaVA configuration but transfer across diverse multimodal model families.

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%
Multimodal PRISM	68.2	65.1	48.9	85.3	86.0	83.1	30.2	63.0	1455.0	305.0	34.1	97.8%
✓ **Visual-only PRISM**	**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 9: **Comparison of selection using visual-only features versus joint multimodal features.** The visual-centric approach (PRISM) demonstrates higher stability as text prompts often introduce non-semantic noise.

space itself. Our work is motivated by the insight that by directly diagnosing and correcting this anisotropy, we can unlock a far more efficient and principled path to data selection that operates on the intrinsic, geometrically-sound structure of the data.

G Theoretical Foundation: Anisotropy as a Source of Computational Overhead

This section analyzes how representation anisotropy can make existing data selection methods for MLLMs inefficient in practice. We formalize the concept of representation anisotropy and show how it degrades simple geometric metrics for measuring semantic diversity. We then discuss why, in the absence of geometric correction, many prior methods rely on proxy signals or training dynamics that introduce substantial overhead, motivating our geometry-aware approach.

G.1 The Geometric Pathology of Anisotropy

We begin by formalizing the geometric properties of the feature space. Let $\phi : \mathcal{D} \to \mathbb{R}^d$ be the MLLM's embedding function. The observed anisotropy in the visual feature space implies a non-zero mean vector $\boldsymbol{\mu} = \mathbb{E}[\phi(d)]$ with a large magnitude. Consequently, any feature vector $\mathbf{x}_i = \phi(d_i)$ can be decomposed into a shared, dominant global component and a sample-specific semantic residual:

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\delta}_i \tag{8}$$

where $\boldsymbol{\delta}_i$ encodes the unique semantics of sample d_i. The defining characteristic of a highly anisotropic space is that this shared component dominates the representation, i.e., $\|\boldsymbol{\mu}\|_2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2]$. This property has a catastrophic consequence for standard geometric metrics, which we formalize in the following theorem.

Theorem 3 (Corruption of Geometric Proximity)**.** *In an anisotropic representation space where $\|\boldsymbol{\mu}\|_2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2]$, the cosine similarity between any two randomly sampled vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is dominated by the global drift vector $\boldsymbol{\mu}$, masking their true semantic dissimilarity, which is encoded in the residuals $\boldsymbol{\delta}_i, \boldsymbol{\delta}_j$.*

Proof. The cosine similarity is given by:

$$\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{(\boldsymbol{\mu} + \boldsymbol{\delta}_i)^\top (\boldsymbol{\mu} + \boldsymbol{\delta}_j)}{\|\boldsymbol{\mu} + \boldsymbol{\delta}_i\|_2 \|\boldsymbol{\mu} + \boldsymbol{\delta}_j\|_2} \tag{9}$$

The numerator expands to $\|\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^\top(\boldsymbol{\delta}_i + \boldsymbol{\delta}_j) + \boldsymbol{\delta}_i^\top \boldsymbol{\delta}_j$. Since $\|\boldsymbol{\mu}\|_2$ is the dominant term, the denominator can be approximated using a first-order Taylor expansion as $\|\boldsymbol{\mu} + \boldsymbol{\delta}_i\|_2 \approx \|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_i}{\|\boldsymbol{\mu}\|_2}$. Substituting these approximations and retaining

only the leading-order terms yields:

$$\cos(\mathbf{x}_i, \mathbf{x}_j) \approx \frac{\|\boldsymbol{\mu}\|_2^2}{\left(\|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_i}{\|\boldsymbol{\mu}\|_2}\right)\left(\|\boldsymbol{\mu}\|_2 + \frac{\boldsymbol{\mu}^\top \boldsymbol{\delta}_j}{\|\boldsymbol{\mu}\|_2}\right)}$$
$$\approx 1 - \frac{1}{2}\left\|\frac{\boldsymbol{\delta}'_i}{\|\boldsymbol{\mu}\|_2} - \frac{\boldsymbol{\delta}'_j}{\|\boldsymbol{\mu}\|_2}\right\|_2^2 \approx 1 \tag{10}$$

where $\boldsymbol{\delta}'$ is the component of $\boldsymbol{\delta}$ orthogonal to $\boldsymbol{\mu}$. The true semantic similarity, proportional to $\boldsymbol{\delta}_i^\top \boldsymbol{\delta}_j$, is relegated to lower-order terms that are numerically insignificant, rendering the metric incapable of distinguishing unique samples. □

G.2 Practical Cost of Ignoring Anisotropy

The failure of simple geometric proxies, as established in Theorem 3, creates a practical dilemma for data selection. When raw feature distances are dominated by global drift, reliable semantic selection typically requires either correcting the geometry or introducing auxiliary signals to approximate the latent semantic structure of the data. This practical cost account explains why many existing visual-instruction selection pipelines incur substantial overhead.

Common practical strategies include:

1. **Implicit Structure Recovery:** Employing iterative clustering or concept-discovery algorithms to recover semantic groups in a distorted space. Such procedures typically introduce costs such as $\mathcal{O}(I \cdot N \cdot K \cdot d)$, where I is the number of iterations and K is the number of clusters.

2. **Bypassing Geometry via Training Dynamics:** Using model gradients or training loss as a proxy for data value. These signals can be effective but require forward and/or backward passes over the data, with costs on the order of $\mathcal{O}(N \cdot C_{\text{pass}})$, where C_{pass} is the cost of a model pass.

Methods such as GraNd (Paul et al., 2023) and TIVE (Liu et al., 2024d) exemplify the second pattern: they obtain useful value estimates from training dynamics, but the resulting selection stage can become comparable to, or even exceed, the cost saved by later training on a smaller subset. PRISM takes a different route. By removing the dominant first-order drift before scoring redundancy, it corrects the usefulness of inexpensive correlation-based geometry and avoids these additional proxy or training-dynamics costs.

H Cross-Architecture Diagnosis of Representation Anisotropy

To further validate that representation anisotropy is a universal characteristic of Multimodal Large Language Models (MLLMs), we expand our diagnostic analysis across five distinct model-encoder combinations. This section provides detailed statistical distributions of per-dimension means to characterize the geometric shift in the feature space.

H.1 Metric Definitions

We utilize the following statistics to quantify the distribution of per-dimension means across different datasets:

- **Min / Max**: The extreme values of the per-dimension mean vectors.
- **Mean($|x|$)**: The arithmetic mean of the absolute per-dimension means, serving as a primary indicator of the magnitude of the global shift from the origin.
- **Percentiles (P25, P75, P99)**: Values representing the spread and skewness of the mean distribution across dimensions.

H.2 Per-dimension Mean Statistics across Configurations

Tables 10 through 13 present the statistics for diverse language backbones and vision encoders. Across all configurations, a consistent pattern emerges: textual features remain well-centered around zero (Mean($|x|$) $\approx$ 0.11–0.13), whereas all visual instruction datasets exhibit substantial non-zero shifts (Mean($|x|$) $\approx$ 0.36–0.66). In an ideal isotropic space, such as the space induced by high-quality text embeddings, the global mean vector should be negligible ($\|\boldsymbol{\mu}\|_2 \approx 0$), and the representation should be dominated by sample-specific variance components $\boldsymbol{\delta}_i$. However, current visual representations from CLIP- and SigLIP-based encoders deviate significantly from this ideal. The per-dimension mean statistics reported here, together with the singular value distributions in Figure 2, show a consistent and substantial non-zero mean shift in visual features. This empirical dominance of the global component directly motivates the condition $\|\boldsymbol{\mu}\|_2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2]$ in Theorem 1,

modeling the observed geometry of current visual encoders rather than an arbitrary theoretical convenience.

H.3 Dataset-Mean Centering Sanity Check

To verify that the observed shift is not merely a small non-zero mean, we further analyze OCRVQA layer-1 features with $N = 500$ samples and $D = 4096$ dimensions. The raw corpus mean has norm $\|\boldsymbol{\mu}\|_2 = 50.8456$, while dataset-mean centering reduces it to 1.8645×10^{-6}. More importantly, the mean component accounts for 0.9721 of the origin second moment:

$$\frac{\|\boldsymbol{\mu}\|_2^2}{\mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top])} = 0.9721. \tag{11}$$

Since $\mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top]) = \|\boldsymbol{\mu}\|_2^2 + \mathbb{E}[\|\mathbf{x} - \boldsymbol{\mu}\|_2^2]$, this implies

$$\frac{\|\boldsymbol{\mu}\|_2^2}{\mathbb{E}[\|\boldsymbol{\delta}\|_2^2]} \approx \frac{0.9721}{0.0279} \approx 34.8, \qquad \frac{\|\boldsymbol{\mu}\|_2}{\sqrt{\mathbb{E}[\|\boldsymbol{\delta}\|_2^2]}} \approx 5.9. \tag{12}$$

Thus, the mean term is not only non-zero but dominates the sample-specific residual energy, directly supporting the assumption $\|\boldsymbol{\mu}\|_2^2 \gg \mathbb{E}[\|\boldsymbol{\delta}_i\|_2^2]$ used in our analysis.

Table 10: **Statistics for Vicuna v1.5 7B + CLIP ViT-L/14 336px.**

Dataset	Min	P25	P75	P99	Max	Mean($\lvert x \rvert$)
Text	-2.8851	-0.1033	0.0064	0.1105	0.3956	0.1310
A-OKVQA	-4.0186	-0.3351	0.0052	0.3329	1.2545	0.4045
GQA	-4.0984	-0.3448	0.0057	0.3404	1.2815	0.4141
LLaVA	-4.0885	-0.3427	0.0046	0.3394	1.2819	0.4126
OCRVQA	-5.6103	-0.5129	0.0023	0.5187	1.8318	0.6127
VG	-4.1034	-0.3449	0.0062	0.3413	1.2827	0.4147

Table 11: **Statistics for Phi-2 2.7B + SigLIP-SO400M-Patch14-384.**

Dataset	Min	P25	P75	P99	Max	Mean($\lvert x \rvert$)
Text	-2.9124	-0.1047	0.0072	0.1119	0.4012	0.1328
A-OKVQA	-4.3215	-0.3587	0.0048	0.3612	1.3421	0.4384
GQA	-4.4102	-0.3701	0.0053	0.3689	1.3745	0.4501
LLaVA	-4.4018	-0.3675	0.0042	0.3668	1.3751	0.4483
OCRVQA	-6.0321	-0.5514	0.0019	0.5589	1.9647	0.6589
VG	-4.4147	-0.3704	0.0058	0.3695	1.3754	0.4505

Table 12: **Statistics for Qwen2.5 7B Base + CLIP ViT-L/14 336px.**

Dataset	Min	P25	P75	P99	Max	Mean($\lvert x \rvert$)
Text	-2.6734	-0.0891	0.0058	0.0945	0.3512	0.1089
A-OKVQA	-3.7845	-0.3012	0.0047	0.3089	1.1523	0.3698
GQA	-3.8621	-0.3101	0.0052	0.3154	1.1789	0.3789
LLaVA	-3.8534	-0.3079	0.0041	0.3147	1.1793	0.3776
OCRVQA	-5.1987	-0.4687	0.0021	0.4823	1.6845	0.5634
VG	-3.8665	-0.3104	0.0056	0.3162	1.1795	0.3793

Table 13: **Statistics for Llama-3-8B + CLIP ViT-L/14 336px.**

Dataset	Min	P25	P75	P99	Max	Mean($\lvert x \rvert$)
Text	-2.6512	-0.0876	0.0055	0.0929	0.3445	0.1067
A-OKVQA	-3.7621	-0.2989	0.0046	0.3067	1.1434	0.3667
GQA	-3.8389	-0.3078	0.0051	0.3132	1.1701	0.3756
LLaVA	-3.8301	-0.3056	0.0040	0.3125	1.1705	0.3743
OCRVQA	-5.1623	-0.4634	0.0020	0.4789	1.6712	0.5589
VG	-3.8432	-0.3081	0.0055	0.3140	1.1707	0.3760

H.4 Architectural Generality and Observations

The results demonstrate that representation anisotropy is a robust and persistent phenomenon across different language backbones (Vicuna, Phi-2, Qwen, Llama) and vision encoders (CLIP, SigLIP). While more advanced language models slightly reduce the absolute magnitude of visual anisotropy (e.g., Llama-3 and Qwen show approximately 10% lower Mean($|x|$) than Vicuna), the visual-textual gap remains pronounced in all cases. This evidence confirms that the geometric origins of Global Semantic Drift are general to the MLLM architecture and are not specific to the LLaVA framework.

I Benchmark Details

We provide detailed descriptions of the seven multimodal benchmarks used in our main evaluation.

MMBench (Liu et al., 2024c) is a multi-faceted benchmark that evaluates the all-around capabilities of MLLMs. It assesses both perception and reasoning through single-choice questions organized hierarchically across a wide range of domains.

ScienceQA (Lu et al., 2022) evaluates scientific question answering. It features multimodal questions from various science domains that require models to jointly reason over visual and textual contexts, often involving diagrams, charts, and complex text.

MME (Fu et al., 2025) is a comprehensive benchmark designed to assess multimodal reasoning and perception skills. It comprises 14 sub-tasks that cover a broad spectrum of capabilities, from object existence and position to text translation and code reasoning.

POPE (Li et al., 2023) is an evaluation framework specifically designed to measure object hallucination in MLLMs. It uses a polling-based (Yes/No) question-answering approach to probe whether a model falsely perceives objects that are not present in an image.

VizWiz (Gurari et al., 2018) tests a model's ability to answer visual questions taken "in-the-wild" by visually impaired users. It serves as a challenging test for zero-shot generalization, as the questions and images are often unconventional and noisy.

MM-Vet (Yu et al., 2024b) is an evaluation benchmark designed to measure the emergent, instruction-following capabilities of MLLMs in a multimodal context. It assesses a model's ability to handle complex prompts that integrate perception, reasoning, and world knowledge.

MMMU (Yue et al., 2023) is a massive, multi-discipline benchmark that assesses expert-level MLLM capabilities. It requires the integration of perception, knowledge, and reasoning across a wide array of college-level subjects, from science and engineering to arts and humanities.

J Sensitivity Analysis of the Selection Ratio

In this section, we provide a detailed sensitivity analysis of the selection ratio τ, which governs the proportion of data retained for instruction tuning. Unlike static hyperparameters, τ serves as a flexible selection budget that practitioners can adjust based on specific computational constraints. Table 14 presents the performance trade-offs across various sampling ratios. Our analysis reveals that PRISM is robust across diverse selection budgets:

- **Monotonic Visual Scaling**: Multimodal performance scales consistently with the data budget, with the 30% ratio (PRISM-Instruct-250K) effectively surpassing the full-dataset baseline.
- **Language Capability Preservation**: Textual performance remains remarkably high even at extreme pruning levels (e.g., 5%), suggesting that PRISM's redundancy identification excels at isolating a high-quality "data core" that minimizes catastrophic forgetting.

K Selection Metric Rationale: Correlation vs. Second-Order Distance

The design of PRISM prioritizes a training-free, computationally efficient pipeline. While second-order measures like Mahalanobis distance are theoretically relevant to anisotropy, they introduce significant bottlenecks in large-scale multimodal settings. We justify the use of implicit re-centering via Pearson-style correlation based on the following factors:

Sampling Ratio (τ)	Visual Performance	Textual Performance
5%	92.3	104.4
10%	95.1	102.9
20%	97.1	102.1
✔ **30% (PRISM Default)**	**101.7**	**101.9**
Full Fine-tuning (Baseline)	100.0	100.0

Table 14: **Performance robustness across varied selection ratios τ.** Visual performance denotes the relative average score across multimodal benchmarks, while textual performance reflects foundational language understanding scores relative to the baseline.

K.1 Algorithmic Efficiency and Scalability

The estimation of Mahalanobis distance requires the computation and inversion of a covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For modern MLLMs with high-dimensional feature spaces, this incurs an $\mathcal{O}(Nd^2)$ cost for estimation and $\mathcal{O}(d^3)$ for inversion. In contrast, PRISM uses corpus-mean re-centering and an exact aggregate form of Pearson-style normalized inner-product scoring, giving an $\mathcal{O}(Nd)$ scoring cost and enabling the processing of the LLaVA-665K dataset in just 1.5 GPU hours.

K.2 Numerical Stability in Anisotropic Spaces

Our singular value analysis demonstrates that pre-trained visual features occupy a low effective rank, where variance is concentrated in a few dominant dimensions. This geometric degeneration often results in ill-conditioned covariance matrices. Relying on matrix inversion in such spaces would require additional regularization (e.g., eigenvalue clipping or shrinkage), introducing new hyperparameters and compromising the simplicity of the training-free paradigm.

K.3 Complexity of Correlation Computation

Although the redundancy score is written as an average over correlations, PRISM evaluates it through an exact aggregate form rather than enumerating all sample pairs. After global re-centering and normalization, each correlation is an inner product between transformed features, so the average correlation of every sample can be obtained from a single global normalized-feature mean. This yields an $\mathcal{O}(Nd)$ scoring procedure and avoids materializing any $N \times N$ matrix. The full derivation, numerical-stability details, and chunked GPU implementation are provided in Appendix M.

L Qualitative Analysis

To provide a more intuitive understanding of the Intrinsic Selection paradigm, we conduct a qualitative analysis of samples ranked by PRISM. Figure 9 illustrates representative examples from the top-ranked (selected) and bottom-ranked (pruned) tiers of the LLaVA-665K dataset.

L.1 High-Value: Challenging and Informative

The top-ranked samples prioritized by PRISM typically involve **fine-grained perception** and **complex contextual reasoning**. For instance, in the "Asia street" example (Figure 9, Left), the model must identify localized visual cues such as the road marking Tomare, the specific structure of utility poles, and architectural styles to correctly determine the geographical region. From a training perspective, these informative interactions provide a stronger learning signal, often corresponding to higher initial loss values, which is critical for mastering expert-level reasoning.

L.2 Low-Value: Trivial and Redundant

Conversely, bottom-ranked samples are frequently characterized by **trivial perception** or **high semantic redundancy**. In the "Sunny weather" example (Figure 9, Right), the task involves identifying a generic weather condition in a common park setting. Such visual scenes—comprising grass, trees, and sunlight—are over-represented in the corpus, leading to a high correlation with the Global Semantic Drift.

Because these samples offer saturated information and redundant gradients, training on them yields diminishing returns and may even exacerbate "knowledge forgetting". By pruning these trivial samples, PRISM ensures a more efficient use of the computational budget without sacrificing model performance.

M Detailed Runtime Analysis

M.1 Wall-Clock Runtime Breakdown

To provide a transparent account of the computational costs associated with different selection methods, we present a detailed breakdown of the wall-clock runtime in Table 15. All experiments were conducted on an NVIDIA A100 80GB GPU, and time was measured in hours.

The table decomposes the total pipeline duration into two key phases: **Data Selection** (the cost of the selection method itself) and **Visual Instruction Tuning** (the cost of fine-tuning the model on the selected data subset).

- **Full-Finetune:** This serves as our baseline, requiring 94 GPU hours for tuning on the complete dataset, with zero selection cost.
- **TIVE:** This Training-Based method exemplifies the high overhead of existing approaches. The selection phase alone consumes 87 hours, which is 92% of the baseline's entire tuning time. Although the final tuning is fast (14 hours), the total pipeline (101 hours) is 7.5% slower than simply fine-tuning on all data, resulting in an OSC score greater than 1.
- **PRISM:** Our method demonstrates exceptional efficiency. The entire selection process, including feature extraction and correlation computation, takes only 1.5 hours. This allows for a total pipeline time of just 29.5 hours, a **71% reduction** compared to the full fine-tuning baseline, robustly satisfying our efficiency principles.

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 15: Wall-clock runtime breakdown on an NVIDIA A100 80GB GPU. All times are measured in hours. Percentages in the 'Overall' column indicate the change relative to the Full-Finetune baseline.

M.2 Pairwise Redundancy Objective and Linear-Time Computation

A natural concern is that the redundancy score in Eq. 6 appears to require all pairwise correlations. PRISM's conceptual objective is the full pairwise redundancy objective over the re-centered corpus. Let $K_{ij} = \rho(\mathbf{F}_i, \mathbf{F}_j)$ denote the centered cosine affinity between two samples. Then $\mathcal{R}(d_i) = \frac{1}{N-1}\sum_{j\neq i} K_{ij}$ is the normalized weighted degree of d_i in the complete corpus-level similarity graph, equivalently a kernel-density estimate in the re-centered feature space. A high value means that the sample lies in a dense, repeatedly covered semantic region; a low value means that it has weak affinity to the dominant redundant modes and therefore increases the selected subset's coverage of less duplicated visual semantics. This is why selecting the low-correlation tail directly targets novelty,

diversity, and reduced over-representation in the instruction pool.

The aggregate expression below provides an algebraic acceleration of this complete-graph score. Define the centered and normalized feature

$$\mathbf{g}_i = \frac{\mathbf{F}_i - \boldsymbol{\mu}_{\mathcal{F}}}{\|\mathbf{F}_i - \boldsymbol{\mu}_{\mathcal{F}}\|_2}, \tag{13}$$

so that $\rho(\mathbf{F}_i, \mathbf{F}_j) = \mathbf{g}_i^\top \mathbf{g}_j$. Let $\bar{\mathbf{g}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{g}_j$ be the sufficient statistic that sums all normalized centered features. Then the same leave-one-out pairwise average can be evaluated as

$$\begin{aligned}\mathcal{R}(d_i) &= \frac{1}{N-1}\sum_{j\neq i} \mathbf{g}_i^\top \mathbf{g}_j \\ &= \frac{\mathbf{g}_i^\top \sum_{j=1}^{N} \mathbf{g}_j - \mathbf{g}_i^\top \mathbf{g}_i}{N-1} \\ &= \frac{N\mathbf{g}_i^\top \bar{\mathbf{g}} - 1}{N-1},\end{aligned} \tag{14}$$

where the last equality follows from $\|\mathbf{g}_i\|_2 = 1$. The vector $\bar{\mathbf{g}}$ is a sufficient statistic that compresses the complete pairwise graph under the linear centered-cosine kernel. It preserves the exact ranking induced by the full average-correlation score while avoiding explicit enumeration of all pairs. Once $\boldsymbol{\mu}_{\mathcal{F}}$ and this sufficient statistic are computed, the N complete-graph degree scores can be obtained in linear time.

In implementation, PRISM processes the feature matrix in chunks. The first streaming pass accumulates the corpus-level mean $\boldsymbol{\mu}_{\mathcal{F}}$. The second pass centers and normalizes each chunk, accumulates $\bar{\mathbf{g}}$, and optionally writes the normalized chunk to disk. The final pass streams stored or recomputed normalized chunks, computes $\mathbf{g}_i^\top \bar{\mathbf{g}}$ for each sample, and writes the redundancy score. At no point do we materialize the $N \times N$ correlation matrix or any block of pairwise correlations.

For numerical stability, the normalization uses a small ϵ in the denominator for near-zero centered vectors. In this case, the implementation subtracts the actual self term $\mathbf{g}_i^\top \mathbf{g}_i$ in the leave-one-out formula; when vectors are exactly unit-normalized, this term reduces to 1. Reductions are accumulated in FP32, and GPU kernels perform the chunk-wise normalization, reductions, and dot products. CPU work is limited to feature loading, chunk scheduling, and writing scores.

For N samples with feature dimension d, the score computation costs $\mathcal{O}(Nd)$ after feature extraction. With chunk size B, the active memory is $\mathcal{O}(Bd + d + N)$ if all scores are kept in memory, or $\mathcal{O}(Bd + d)$ when scores are streamed to disk. This aggregate streaming implementation explains why processing LLaVA-665K takes only 1.5 hours on an A100 80GB GPU, far below the 94 hours required for full fine-tuning and the 87 hours required by the selection stage of TIVE.

Interpretation of the redundancy score. PRISM's redundancy score should be interpreted as a corpus-level average-correlation degree under the centered-cosine kernel, rather than as a local nearest-neighbor density estimator. This design is intentional. Local density methods such as kNN-based scoring can capture fine-grained neighborhoods, but they require either approximate nearest-neighbor search or explicit pairwise comparisons at scale. PRISM instead estimates whether a sample aligns with the dominant centered semantic mass of the corpus. Samples with low average correlation are less explained by this corpus-level semantic mass and are therefore more likely to contribute complementary information during instruction tuning.

Outliers versus useful diversity. A potential concern is that low-correlation samples may include outliers. In practice, instruction-tuning datasets are already curated to contain valid image-instruction pairs, and PRISM is applied within this curated pool. We therefore interpret low correlation not as arbitrary outlierness, but as low redundancy relative to the instruction corpus. The qualitative examples in Appendix K further show that high-ranked samples tend to contain more specific visual reasoning, OCR, or compositional information, whereas low-ranked samples are often visually or semantically repetitive.