DICS: Exploring Data Intrinsic Consistency for Visual Instruction Selection

Anonymous ACL submission

Abstract

Visual instruction tuning is crucial for advancing the vision-language alignment and instruction-following capabilities of Vision-Language Models (VLMs). However, identifying optimal subsets under a fixed ratio constraint from rapidly expanding datasets remains a significant bottleneck. While existing methods largely depend on distribution diversity or heuristic filtering, they often overlook the internal coherence within individual samples. To bridge this gap, we propose *Data Intrinsic Consistency (DIC)*, a self-scoring metric designed to quantify the sample-level inter-component consistency. DIC consists of two modules: *Visual Information Consistency (VIC)*, evaluating the alignment between visual content and instructions, and *Response Information Consistency (RIC)*, assessing response coherence relative to the instruction. Building upon DIC, we introduce *Data Intrinsic Consistency Selection (DICS)*, an adaptive data selection method that optimizes the trade-off between high intra-sample consistency and global distributional diversity under varying data budgets. Extensive experiments demonstrate that DICS consistently outperforms state-of-the-art methods across diverse dataset scales and model architectures, surpassing full-dataset fine-tuning while using only 25% of the LLaVA-1.5-665K data. We further curate DICS-6M, a 6M-sample multi-modal instruction corpus that enables the largest-scale visual instruction selection study to date; remarkably, DICS reaches 94.52% of the official InternVL3-8B-Instruct performance using less than 25% of its reported training data. Curated datasets, source code, and trained models will be publicly released.

1 Introduction

Visual instruction tuning (Ouyang et al., 2022; Cui et al., 2023; Chiang et al., 2023) is pivotal for VLM training, establishing foundational capabilities while enabling effective human instruction understanding and following. Given the massive scale of such data, rigorous evaluation and selection are critical.

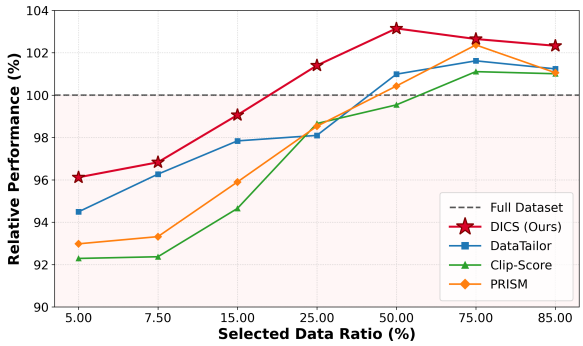

Figure 1: **Performance of LLaVA-1.5-7B fine-tuned on subsets of LLaVA-1.5-665K selected by various methods.** DICS-selected subsets outperform the full dataset at 25% and consistently surpass other methods.

Existing selection methods are mainly based on distribution diversity (Bi et al., 2025), sample relationships (Yu et al., 2025), importance (Wu et al., 2024), and heuristic quality rules (Saada et al., 2025; Li et al., 2023b) to identify compact subsets, removing redundancy to achieve superior VLM performance. However, these methods neglect the intrinsic relationships among images, instructions, and responses within individual samples. While similar efforts exist for language data (Li et al., 2024b), they are inapplicable to multimodal settings due to the lack of mechanisms to evaluate visual information's impact on quality.

In this work, we propose **Data Intrinsic Consistency (DIC)**, a sample-level self-scoring metric grounded in the hypothesis that high-quality visual instruction data must exhibit internal consistency among its image, instruction, and response components. DIC consists of two modules: **Visual Information Consistency (VIC)**, which assesses whether the image is directly relevant to the instruction-response pair and provides critical context for generating the response; and **Re-**

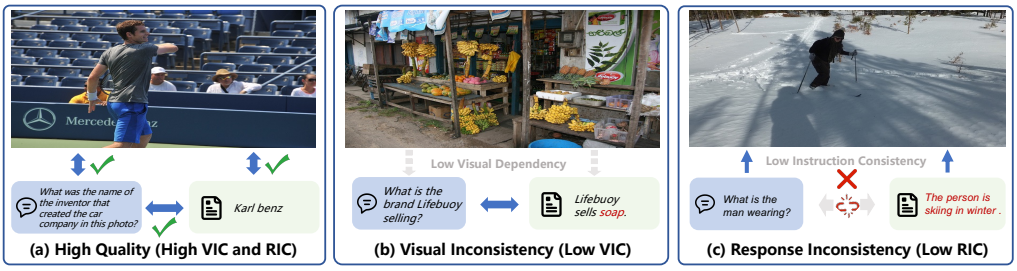

Figure 2: **Examples of Data Intrinsic Consistency in Visual Instruction Data.** (a) High-quality samples exhibiting both high VIC and RIC. (b) Low-VIC samples showing low visual dependency between cross-modal inputs and responses. (c) Low-RIC samples showing weak alignment of the response with the image and instruction.

sponse Information Consistency (RIC), which evaluates whether the response is focused on the instruction, accurate, and logically coherent given the visual input. Fig. 2 illustrates examples of inconsistency between images, instructions, and responses in various visual instruction samples. Intuitively, low-VIC samples exhibit weak alignment between instructions and images, making them prone to hallucinations. Similarly, low-RIC samples provide responses clearly irrelevant to both the instruction and visual content, potentially disrupting the model's vision-language alignment. To unify VIC and RIC, we introduce the DIC metric via a self-scoring framework that contrasts training losses under partial versus complete instruction scenarios. A larger loss discrepancy signifies a tighter cross-modal alignment and richer learning potential, effectively filtering high-quality samples to enhance accuracy and mitigate hallucinations.

Correspondingly, we introduce **Data Intrinsic Consistency Selection (DICS)**, a data selection method that jointly considers DIC scores, data diversity, and sampling efficiency under varying budget constraints. Its core principle involves an adaptive sampling strategy: when the selection ratio is low, it prioritizes high-DIC samples while maintaining data diversity. Conversely, when the selection ratio is high, it removes low-DIC samples exhibiting redundant features. This approach ensures the consistent selection of high-quality subsets across different data scales and resource limitations.

Extensive experiments demonstrate the significant efficacy of our DIC-based selection method in visual instruction tuning. As shown in Fig. 1, on the widely adopted LLaVA-1.5-665K dataset, our method consistently outperforms state-of-the-art baselines across all sampling ratios. Notably, a mere 25% subset selected by our approach achieves performance comparable to full-data fine-tuning. Furthermore, as the sampling ratio increases, our method consistently surpasses full-data baselines. To validate effectiveness and practicality on large-scale datasets, we curated **DICS-6M**, a 6M-sample corpus assembled entirely from open-source datasets following the InternVL3 (Zhu et al., 2025) collection protocol, and conducted the largest data selection experiment to date on the open-source InternVL3-8B (Zhu et al., 2025). Our results show that using only 25% of the data yields superior performance compared to fine-tuning on the full 6M dataset. Compared against the more advanced official InternVL3-8B-Instruct, our model matches 94.52% of its performance with less than 25% of its training data (5.1M vs. 21.7M). We will open-source our trained models and datasets. In summary, our contributions are as follows:

- We propose Data Intrinsic Consistency, comprising Visual Information Consistency and Response Information Consistency, which leverages a unified evaluation paradigm to quantify sample-level intrinsic consistency among images, instructions, and responses.
- We design a unified DICS sampling strategy that integrates DIC scores, diversity, and sampling efficiency to achieve budget-aware optimal subset selection.
- Comprehensive experiments demonstrate that DICS-selected datasets achieve SOTA performance on various benchmarks, exhibiting superior robustness and scalability across varying data scales and model architectures.

2 Related Work

2.1 Visual Instruction Tuning

Visual instruction tuning (Ouyang et al., 2022; Cui et al., 2023; Chiang et al., 2023) is essential for aligning Large Language Models (LLMs) with visual modalities. Following the instruction-tuning in

NLP (Lou et al., 2024; Tinn et al., 2023), seminal works like LLaVA (Liu et al., 2023) and MiniGPT-4 (Zhu et al., 2024) utilized GPT-4 generated data to bridge visual encoders with LLMs. The field has since evolved through two main trajectories: **(1) Data Scaling**, exemplified by ShareGPT4V (Chen et al., 2024a) and InstructBLIP (Dai et al., 2023), which emphasize expanding dataset scale and task variety; and **(2) Architectural Refinement**, where models like BLIP-2 (Li et al., 2023a) introduce novel bottleneck structures, and recent models like LLaVA-NeXT (Li et al., 2024a) and Qwen2-VL (Wang et al., 2024) adopt dynamic resolution mechanisms to capture fine-grained visual details across arbitrary scales.

2.2 Visual Instruction Selection

To mitigate the computational costs of training on massive datasets, recent works focus on identifying high-value subsets and generally fall into two categories: representation-based and gradient-based methods. **(1) Representation-based methods**: These approaches involve filtering data based on global distributional properties to reduce redundancy. Methods like DataTailor (Yu et al., 2025) and PRISM (Bi et al., 2025) utilize embedding-based clustering to ensure concept coverage and select representative samples. Other approaches (Tang et al., 2025; Hong et al., 2025) further optimize for distinctiveness within the feature space. Additionally, ARDS (Yang et al., 2025) constructs robust training mixtures by prioritizing samples semantically close to worst-case evaluation subgroups through clustering and perturbations. These methods are effective at maintaining diversity and ensuring the selected subset covers the original distribution. **(2) Gradient-based methods**: These methods assess data value through training dynamics. LESS (Xia et al., 2024) computes low-rank gradient similarity to select data that benefits specific target tasks. Similarly, COINCIDE (Lee et al., 2024), ICONS (Wu et al., 2024) and TIVE (Liu et al., 2024b) employ influence functions to identify samples that consistently contribute to model generalization or task difficulty. These methods focus on the relational influence of each data sample to maximize transferability. Although the aforementioned methods effectively reduce dataset size, they share a common limitation: they primarily focus on global diversity or distributional coverage, while largely neglecting the intrinsic quality of individual samples. Furthermore, while recent works (Chen et al., 2024b; Lee and Hwang, 2026) emphasize image-text consistency, they either rely on expensive external LLMs for sample-level filtering, or only consider unidirectional alignment between modalities, both of which are unscalable for million-level training data selection. In this paper, we propose Data Intrinsic Consistency (DIC) to assess sample-level alignment and integrate it into DICS, a unified framework that prunes inconsistent samples while preserving distributional diversity for robust training.

3 Data Intrinsic Consistency

A visual instruction data sample (I, x, y) consists of three components: image I, instruction x, and response y. It plays a pivotal role in fine-tuning VLMs by establishing alignment between vision and language, thereby enhancing visual instruction following capabilities. A high-quality instruction sample should exhibit semantic coherence and mutual information gain among I, x, and y. Specifically, instruction x and response y must be grounded in the visual content of the image I, ensuring that the generated response requires a substantive understanding of the visual input rather than relying solely on language priors. Furthermore, the response y should provide accurate and rich content in information while maintaining strong alignment with both the instruction x and the image I. To this end, we propose **Data Intrinsic Consistency (DIC)**, which evaluates sample quality from two perspectives: Visual Information Consistency (VIC) and Response Information Consistency (RIC).

3.1 Visual Information Consistency

Since instructions and responses that are semantically inconsistent with visual content can disrupt the vision-language alignment established during pre-training, we draw inspiration from the principle of information gain (Shannon, 1948) and prior work (Li et al., 2024b). We quantify VIC by measuring the reduction in predictive uncertainty after observing the image. Specifically, we compute the difference in the model's loss when predicting the response y with and without the image I:

$$\mathrm{VIC}(I, x, y) = \exp(\mathcal{L}_w(y \mid x) - \mathcal{L}_w(y \mid x, I)), \tag{1}$$

where $\mathcal{L}_w(\cdot)$ denotes the cross-entropy loss weighted by token importance. The $\exp(\cdot)$ maps the loss difference into the positive domain. A

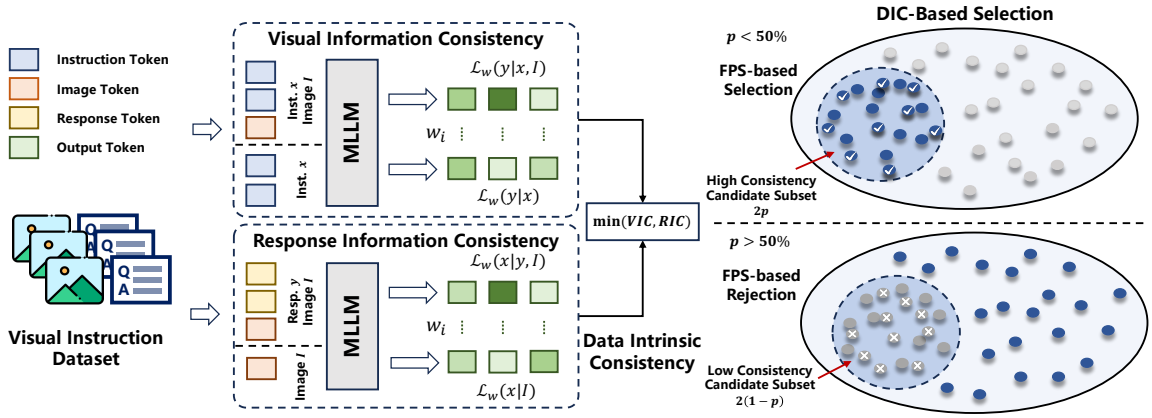

Figure 3: **Overview of Data Intrinsic Consistency Selection (DICS).** First, we compute the Visual Information Consistency and Response Information Consistency for each sample to derive its Data Intrinsic Consistency (DIC) score. Subsequently, leveraging these DIC scores and a target sampling ratio, we employ a designed adaptive sampling strategy to construct a dataset characterized by high intrinsic consistency and preserved diversity.

higher VIC score indicates that the image provides substantial information gain for response generation. To boost semantic sensitivity, we employ a POS-based token weighting strategy to suppress neutral tokens (e.g., prepositions). The weighted loss is defined as:

$$\mathcal{L}_w(y \mid c) = -\frac{\sum_{i=1}^{N} w_i \log p(t_i \mid c, t_{<i})}{\sum_{i=1}^{N} w_i}, \quad (2)$$

where N, t_i, $t_{<i}$, and c denote the token count, current token, preceding context, and conditioning input (x or (x, I)), respectively; and w_i represents the POS-derived weight. Implementation details are in Appendix A.1. Then the VIC score can be expressed as:

$$\mathrm{VIC}(I, x, y) = \exp\left(\frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \Big[-\log p(t_i \mid x, t_{<i}) + \log p(t_i \mid x, t_{<i}, I) \Big]\right). \quad (3)$$

Samples with responses strongly grounded in visual content receive higher VIC scores, whereas vision-irrelevant or hallucinatory samples score lower. For multi-turn dialogues, the VIC score is averaged across all turns.

3.2 Response Information Consistency

To evaluate the consistency between response y and instruction x, we introduce Response Information Consistency (RIC). It assesses if a response is specifically tailored to the query by framing it as an "inverse prediction" task: conditioned on visual context I, we measure how well y aids reconstructing the original instruction x. Specifically, we employ carefully designed prompts (Appendix A.3) to guide the model in inferring the original instruction from the response. For clarity, we omit the explicit prompt tokens and define the RIC as:

$$\mathrm{RIC}(I, x, y) = \exp(\mathcal{L}_w(x \mid I) - \mathcal{L}_w(x \mid y, I)), \quad (4)$$

where $\mathcal{L}_w(\cdot)$ is the token-weighted cross-entropy loss. The difference between $\mathcal{L}_w(x \mid I)$ (predicting the instruction from the image without response) and $\mathcal{L}_w(x \mid y, I)$ (with response) quantifies the information gain provided by y for reconstructing x. Following the same token-weighting strategy, RIC can be expressed as:

$$\mathrm{RIC}(I, x, y) = \exp\left(\frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \Big[-\log p(t_i \mid I, t_{<i}) + \log p(t_i \mid y, I, t_{<i}) \Big]\right). \quad (5)$$

Intuitively, if y aids predicting instruction tokens ($p(t_i \mid y, I, t_{<i}) > p(t_i \mid I, t_{<i})$), it positively contributes to the RIC score. High scores identify samples with coherent semantics alignment. For multi-turn dialogues, we average scores across turns to capture overall consistency.

Data Intrinsic Consistency. For each sample, we define its Data Intrinsic Consistency (DIC) score by taking the minimum of VIC and RIC:

$$\mathrm{DIC}(I, x, y) = \min(\mathrm{VIC}(I, x, y), \mathrm{RIC}(I, x, y)). \quad (6)$$

Eq. 6 enforces that high-quality samples must exhibit strong consistency across *both dimensions*: the image must be informative for generating the response, and the response must be precisely aligned with the instruction.

4 Instruction Data Selection

Given a visual instruction dataset $\mathcal{D}$, our goal is to select a subset $\mathcal{D}^*$ based on DIC scores and a sampling ratio p. The selection process is guided by three key principles: (1) *High DIC Score*: Prioritize samples with high DIC scores to maximize information consistency during training. (2) *Diversity Preservation*: Maintain semantic coverage to prevent distribution shift in the selected subset. (3) *Budget Adaptability*: Balance DIC quality and semantic diversity across sampling ratio p.

To this end, we propose **Data Intrinsic Consistency Selection (DICS)**, an adaptive data selection method based on DIC. Fig. 3 illustrates the data selection pipeline, while Algorithm 1 in appendix shows the algorithmic pseudocode. The selection process comprises two stages: (1) *Scoring and Ranking*: Compute the DIC score for each sample in $\mathcal{D}$ and rank them accordingly. (2) *Adaptive Sampling*: Perform an *adaptive selection strategy* based on the DIC ranking and the target ratio p using the Farthest Point Sampling (FPS) algorithm. Specifically, the *adaptive selection strategy* adapts to the sampling ratio p as follows:

- **Low Ratio** ($p < 50\%$): A candidate pool is formed consisting of the top-$2p$ fraction of the ranked samples. The FPS then selects the fraction p from this pool to balance quality and diversity.
- **Boundary** ($p = 50\%$): Top 50% samples are directly selected.
- **High Ratio** ($p > 50\%$): The top-$(2p-1)$ fraction forms the core set. Afterward, FPS selects an additional $(1-p)$ fraction from the remaining samples to fill the quota.

The core idea behind this selection strategy is twofold: when selecting a small proportion of data, the method prioritizes high-DIC samples while maintaining diversity; conversely, when selecting a large proportion, it removes redundant samples with low DIC scores from the pool. This strategy for large-scale selection mirrors practical filtering approaches, effectively striking a balance between selection efficiency and data quality.

5 Experiments

5.1 Experimental Settings

Datasets & Implementation Details. We conduct experiments on two widely adopted open-source visual instruction datasets: **LLaVA-1.5-665K** and **Vision-FLAN-186K**, utilizing the LLaVA-1.5-7B-Base backbone initialized from LLaVA-1.5-Captions-558K. To further validate the efficacy of DICS at scale, we curate **DICS-6M**, a dataset comprising 6 million high-quality instruction samples collected according to the official InternVL3 open-source strategy (details in Appendix B.1), and fine-tune the InternVL3-8B-Pretrained model on it. Regarding training protocols, we fine-tune LLaVA-1.5 with LoRA on 8 NVIDIA A100 (40GB) GPUs, and fully fine-tune InternVL3 on 32 GPUs, both for one epoch. For DIC computation, the DIC score of text-only samples is fixed to 1.0, since no visual information is present. To ensure stability during feature extraction (initiated from the final token of the last layer), we briefly warm up the base model on a 5% random subset before scoring. The warm-up checkpoint is used only for scoring and is then discarded; all final models are trained from the original pre-trained checkpoint. Unless otherwise specified, all reported results are obtained on 25% subsets of the respective full datasets.

Benchmarks. We assess the capabilities of fine-tuned models across 12 benchmarks spanning five distinct dimensions: (1) General Capability (MME (Fu et al., 2025), MMBench (Liu et al., 2024a)); (2) Knowledge and Reasoning (MMMU (Yue et al., 2024), SQA-I (Lu et al., 2022), VizWiz (Gurari et al., 2018)); (3) Document Understanding (AI2D (Kembhavi et al., 2016), DocVQA (Mathew et al., 2020), InfoVQA (Mathew et al., 2021)); (4) Hallucination Mitigation (POPE (Li et al., 2023c), Hallusion-Bench (Guan et al., 2024)); and (5) Open-ended Dialogue (MM-Vet (Yu et al., 2023)). Further details are provided in the Appendix B.2.

Baselines. We conduct a comparative analysis of DICS against four categories of baseline methods: (1) Stochastic Sampling via random selection to ensure uniform data diversity; (2) Heuristic Filtering based on data length (Zhao et al., 2024) and CLIP-Score (Radford et al., 2021); (3) LLM-based Selectors, such as Perplexity (Marion et al., 2023) and IFD (Li et al., 2024b), which assess data quality via loss differentials within LLMs; and (4) MLLM-based Selectors, including ARDS (Yang et al., 2025), COINCIDE (Lee et al., 2024), DataTailor (Yu et al., 2025), and PRISM (Bi et al., 2025). All baselines follow their official implementations. See Appendix B.3 for details.

5.2 Main Results

Efficacy of DICS. Table 1 presents the results of various data selection methods on the LLaVA-1.5-

Table 1: **Performance comparison of data selection methods using 25% of the LLaVA-1.5-665K dataset.** The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Scale	General Capability			Knowledge & Reasoning			Document Understanding			Hallucination		Dialogue	
		MME	MMBench en	MMBench cn	MMMU	SQA-I	VizWiz	AI2D	DocVQA	InfoVQA	POPE	Hallusion Bench	MMVet	Rel. (%)
Full Dataset	665K	59.92	60.22	51.93	35.89	68.32	31.79	47.83	23.82	22.30	84.49	27.80	41.01	100.00
Random Selection	166K	59.28	56.27	51.93	35.56	**66.98**	29.53	45.08	21.62	21.63	81.58	29.46	38.07	96.69
Length (Zhao et al., 2024)	166K	55.92	49.23	43.27	34.22	60.39	27.21	41.06	18.88	19.69	76.70	**34.01**	35.52	88.79
CLIP-Score (Hessel et al., 2021)	166K	58.52	58.20	51.93	34.33	**66.98**	30.32	**51.59**	22.26	21.34	83.46	26.24	**42.80**	98.66
Perplexity (Marion et al., 2023)	166K	54.64	43.19	41.95	33.78	57.91	29.68	46.31	21.64	22.50	84.04	26.07	37.34	89.85
IFD (Li et al., 2024b)	166K	**64.81**	34.83	33.82	32.78	57.76	29.27	44.85	19.84	22.19	84.10	26.70	35.46	87.58
ARDS (Yang et al., 2025)	166K	55.81	**60.06**	53.10	35.89	**66.98**	27.21	**51.59**	18.97	21.49	81.29	28.01	35.28	96.46
COINCIDE (Lee et al., 2024)	166K	58.38	52.55	48.14	34.44	66.73	31.32	51.04	23.05	22.46	84.24	28.06	37.66	96.88
DataTailor (Yu et al., 2025)	166K	59.75	56.42	51.01	34.22	65.59	31.25	46.05	22.64	21.42	80.24	28.92	42.16	97.17
PRISM (Bi et al., 2025)	166K	60.03	56.58	51.55	**36.11**	66.63	**31.77**	46.70	23.87	21.87	82.18	26.49	38.58	97.66
DICS (Ours)	166K	63.39	59.52	**55.03**	**36.11**	66.14	31.29	50.52	**24.64**	**23.11**	**84.91**	26.79	41.74	**101.40**

7B model. Remarkably, DICS achieves an average improvement of 1.40% across all 12 benchmarks compared to full-dataset training. This consistent gain across diverse task categories underscores that intrinsic data consistency outweighs mere scale. By leveraging DICS, we effectively filter out suboptimal samples characterized by low visual dependency or low instruction-response consistency. Specifically, our method attains state-of-the-art performance on key benchmarks, including MMBench-CN (55.03), MMMU (36.11), DocVQA (24.64), InfoVQA (23.11), and POPE (84.91). In contrast, prior heuristic methods often trade generalization for peak performance on specific metrics. For instance, length-based selection maximizes HallusionBench scores by constraining response length, yet this reduces information density and reasoning depth, resulting in a significant drop in overall capability (88.79% relative performance). DICS overcomes these limitations by evaluating intrinsic consistency among images, instructions, and responses, thereby ensuring balanced and robust performance across all dimensions.

Performance across Sampling Ratios. As illustrated in Fig. 1, DICS consistently outperforms representative strong baselines across varying sampling ratios. Under low proportions ($p < 25\%$), DICS demonstrates a distinct advantage, validating that our consistency metrics effectively identify high-value samples. As p increases, performance initially peaks (reaching 103.15% at 50%) before gradually declining and saturating. This occurs because higher sampling ratios incorporate more low-DIC samples, compromising model alignment and instruction-following capabilities. Notably, even after peaking, DICS maintains a lead over competitors. In contrast, PRISM exhibits marked degradation due to the inclusion of low-quality samples. These curves indicate that DICS rapidly identifies beneficial samples while delaying the introduction of noise, thereby maximizing data efficiency. It is worth noting that the optimal sampling rate p is not fixed but depends on specific metrics and datasets. Similarly, other methods exhibit comparable peaks; for instance, both PRISM and DataTailor reach their optimal performance around 75%.

Table 2: **Performance comparison of data selection methods using 25% of Vision-FLAN dataset.**

Method (*Sampling Ratio: 25%*)	General Avg.	Know. Avg.	Doc. Avg.	Hall. Avg.	Dial. Avg.	Rel. (%)
Full Dataset	51.93	42.07	29.54	58.21	35.28	100.00
Random Selection	47.90	41.45	28.34	57.87	34.40	96.34
Length (Zhao et al., 2024)	45.27	38.90	24.55	57.75	36.10	91.47
CLIP-Score (Hessel et al., 2021)	44.46	41.73	33.42	57.74	37.06	97.90
Perplexity (Marion et al., 2023)	30.38	38.28	27.93	56.51	33.35	83.50
IFD (Li et al., 2024b)	44.13	39.63	28.92	55.95	36.38	93.11
DataTailor (Yu et al., 2025)	44.64	40.87	30.98	58.26	36.38	96.18
PRISM (Bi et al., 2025)	49.14	42.26	30.28	58.54	32.29	98.48
DICS (Ours)	47.95	42.25	32.53	58.30	37.06	99.91

Table 3: **Generalization across architectures on LLaVA-1.5-665K.** Average performances are reported.

Model	Selector	Avg. Score	Rel.(%)
LLaVA-1.5-7B	Full Dataset	46.28	100.00
	Random Selection	45.24	97.74
	DICS (LLaVA-1.5-7B)	46.93	101.40
	DICS (Qwen2-VL-7B)	46.36	100.16
Qwen2-VL-7B	Full Dataset	70.46	100.00
	Random Selection	70.61	100.21
	DICS (LLaVA-1.5-7B)	71.35	101.25
	DICS (Qwen2-VL-7B)	71.10	100.91
LLaVA-1.5-13B	Full Dataset	47.89	100.00
	Random Selection	47.86	99.93
	DICS (LLaVA-1.5-7B)	48.75	101.79

5.3 Robustness and Scalability

Cross-Dataset Generalization. To demonstrate robustness across diverse domains, we evaluate DICS on Vision-Flan-186K, which features a distribution distinct from LLaVA-1.5-665K (see Table 2). Remarkably, DICS achieves the best overall relative performance: training on just a 25% subset yields

Table 4: **Performance comparison on the 6M large-scale dataset using InternVL3-8B.**

Method	Scale	General	Know.	Doc.	Hall.	Dial.	Rel.
		Avg.	Avg.	Avg.	Avg.	Avg.	(%)
InternVL3-8B-Instruct	21.7M	84.29	78.80	82.94	67.79	70.10	109.08
Full Dataset (100%)	6.0M	77.62	69.42	77.88	65.73	56.19	100.00
DICS (15%)	0.9M	75.16	68.87	77.29	67.58	56.88	99.26
DICS (25%)	1.5M	77.92	71.71	**78.94**	**68.07**	54.50	101.47
DICS (50%)	3.0M	75.97	70.64	78.01	67.57	57.16	100.32
DICS (75%)	4.5M	79.79	**72.94**	78.47	67.11	52.89	101.86
DICS (85%)	5.1M	**80.53**	72.59	78.76	67.70	**59.13**	**103.10**

Table 5: **Ablation study on DIC.**

Variant	General	Know.	Doc.	Hall.	Dial.	Rel.
(Sampling Ratio: 25%)	Avg.	Avg.	Avg.	Avg.	Avg.	(%)
Baselines						
Full Dataset (100%)	57.36	**45.33**	31.32	56.15	41.01	100.00
FPS (Diversity Only)	54.10	43.40	31.86	55.76	39.08	97.02
Single Consistency Metric						
DIC w/ VIC only	59.03	44.86	32.73	55.82	39.54	101.02
DIC w/ RIC only	53.91	43.92	31.45	**56.98**	40.28	97.60
Metric Fusion Strategies						
Sum (VIC+RIC)	56.43	43.80	31.50	55.40	41.56	98.61
Product (VIC×RIC)	56.24	44.53	31.72	56.20	39.17	98.84
Weighted Token Strategy						
DIC w/o weighted token	58.19	43.93	**32.85**	55.92	40.05	100.25
DIC (Ours)	**59.31**	44.51	32.76	55.85	**41.74**	**101.40**

a relative performance of **99.91%**, outperforming all baselines and closely matching full-dataset fine-tuning. These results confirm that DICS generalizes beyond specific data patterns, serving as a robust indicator for selection. Detailed results are listed in Appendix Table 8.

Cross-Architecture Generalization. To validate compatibility across diverse architectures, we extend evaluation to Qwen2-VL-7B under the same settings. As shown in Table 3, DICS-selected subsets (using Qwen2-VL as the selector) consistently outperform the full-dataset baseline, achieving a relative performance of **100.91%**. This indicates that DICS prioritizes intrinsic data consistency independent of model structure. Beyond single-model generalization, we observe robust cross-architectural transferability: data selected by LLaVA-1.5-7B effectively enhances Qwen2-VL-7B (101.25%), while conversely, subsets selected by Qwen also boost the LLaVA model. This mutual improvement confirms that DIC targets universal sample characteristics rather than model-specific patterns. Notably, the LLaVA-1.5-13B model achieves **101.79%** performance using data selected by the smaller 7B variant. This suggests that DIC score computed by smaller models transfer to larger-scale training, allowing lightweight selectors to reduce the cost of larger data curation.

Scalability across Model Architectures. To validate DICS under modern architectures and with large-scale data, we conducted experiments on **DICS-6M** , a 6M-sample corpus curated following InternVL3's open-source data collection protocol (details in Appendix B.1). We evaluated DICS by comparing our results against models fine-tuned with varying sampling ratios as well as the official InternVL3-8B-Instruct with DICS-6M. As shown in Table 4, using the full 6M dataset as a baseline, DICS consistently outperforms across all sampling ratios. Notably, at sampling ratios exceeding 25%, the fine-tuned model surpasses the performance of the full dataset. Specifically, at an 85% sampling ratio, the model achieves optimal performance, outperforming the full-dataset baseline by 3.10%. Remarkably, despite utilizing less than 25% of the official data volume (5.1M vs. 21.7M), our model demonstrates performance comparable to the official InternVL3-8B-Instruct, achieving a relative score of 94.52% (our 103.10% vs. official 109.08%). These findings underscore two key insights: (1) the superior quality of DICS-6M and the significant redundancy inherent in the original visual instruction data; and (2) the robust scalability and practical utility of DICS.

5.4 Ablation Studies and Analysis

Ablation Studies on DIC. To assess individual contributions of our metrics, we conducted ablation studies on a 25% (166K) subset of LLaVA-1.5-665K (Table 5). Pure diversity-based sampling (FPS, 97.02%) underscores the necessity of quality-aware filtering. Among single metrics, VIC drives significant gains (101.02%) while RIC specifically mitigates hallucinations. However, simple fusions (Sum/Product) fail (∼98%) because high scores in one metric can compensate for low scores in the other, introducing noise. In contrast, our Min-based DIC requires both dimensions to be high. Additionally, removing our weighted token strategy causes a drop from 101.40% to 100.25%, validating its importance. Ultimately, our proposed DIC achieves the peak relative performance (101.40%) by combining visual alignment with logical coherence.

Ablation Study on Selection Strategies. To evaluate sampler impacts, we compared Top-K, Hierarchical Clustering, and Farthest Point Selection (FPS). As shown in Table 6, FPS yielded the best relative performance (**101.40%**) at a sampling ratio of 25%; therefore, we adopt it as our default. In particular, DICS consistently outperforms across

Table 6: **Robustness across selection strategies and combinations.** We evaluate different scoring metrics combined with various sampling methods at a 25% ratio.

Method & Strategy	General	Know.	Doc.	Hall.	Dial.	Rel.
(Sampling Ratio: 25%)	Avg.	Avg.	Avg.	Avg.	Avg.	(%)
Full Dataset	57.36	**45.33**	31.32	56.15	41.01	100.00
DataTailor (Top-K)	56.08	44.07	31.82	55.54	41.51	98.76
DataTailor (Hierarchical)	55.73	43.69	31.76	54.58	**42.16**	98.11
DataTailor (FPS)	54.37	43.86	32.23	53.92	41.51	97.37
PRISM (Top-K)	56.05	44.84	32.43	54.34	38.58	98.53
PRISM (FPS)	54.95	43.86	32.06	55.32	40.23	97.86
DIC (Top-K)	57.04	44.78	**33.06**	55.85	41.28	100.40
DIC (Hierarchical)	58.36	45.28	32.61	**56.60**	41.19	101.39
DIC (FPS, Ours)	**59.31**	44.51	32.76	55.85	41.74	**101.40**

all strategies, whereas DataTailor and PRISM fail to match the full-dataset baseline. This shows that DIC more accurately captures the intrinsic learning value, reliably identifying high-quality samples regardless of the selection method employed.

Visualization of VIC and RIC. Fig. 4 illustrates visual instruction samples with varying VIC and RIC scores. High-VIC samples are characterized by the need to rely on explicit visual cues in the image (e.g., text on logos or book covers) to answer questions; in contrast, low-VIC samples tend to be vision-irrelevant (relying solely on linguistic information) or contain hallucinated content. Complementarily, high-RIC responses exhibit strong semantic consistency and logical coherence, while low-RIC responses often suffer from disconnection and fail to capture the intended meaning. Consequently, samples scoring high in both VIC and RIC represent the ideal core instructional data.

Qualitative Comparison. Fig. 5 visualizes text-to-image attention maps (response tokens to visual patches) for different models. Compared to other methods, the DICS-trained model exhibits a more focused attention distribution that precisely localizes the queried object (e.g., the horse's body), indicating superior visual grounding. More details can be seen in Appendix C.3.

Computational Efficiency. DICS is highly efficient, requiring only inference-time forward passes without gradient updates. On the LLaVA-1.5-665K dataset using 8 A100 GPUs, data selection took ∼12 hours; selecting a 25% subset reduced total training time to 14 hours (a 1-hour saving vs. full-dataset training). In the DICS-6M experiment with 32 A100 GPUs, selection and training were completed in 85 hours, yielding a 5-hour savings over the baseline while achieving a 3.10% performance gain. Since DIC scores are computed per sample, computational cost scales linearly with data volume. This characteristic, combined with its inherent parallelizability, makes DICS increasingly advantageous for large-scale datasets. Detailed cost breakdowns are provided in Appendix C.4.

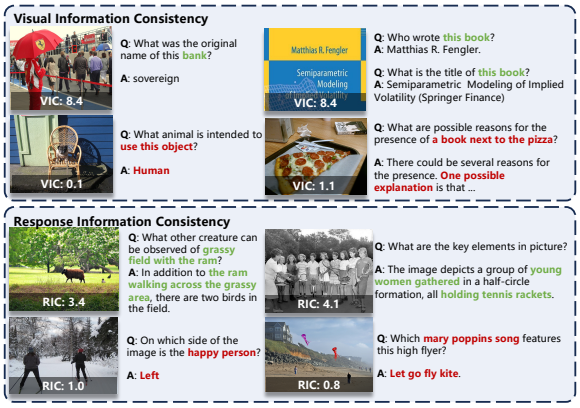

Figure 4: **Visualization of DIC**. We contrast samples with high and low VIC/RIC scores.

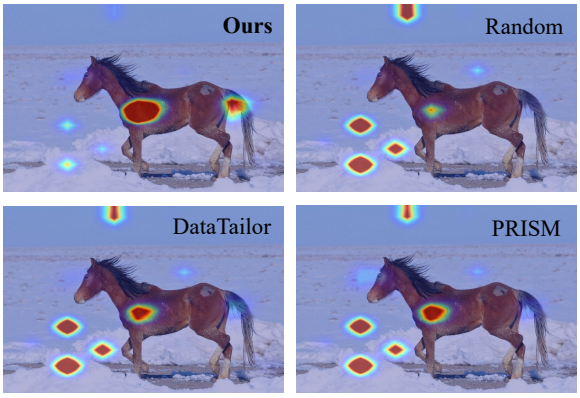

Figure 5: **Qualitative comparison on text-to-image attention maps** for the query: *"What color is the horse in the image?"* (Answer: *"Brown."*). DICS exhibits more precise visual grounding.

6 Conclusion

This paper introduces Data Intrinsic Consistency (DIC), a self-scoring metric quantifying sample-level inter-component consistency. DIC comprises two modules: *Visual Information Consistency (VIC)*, evaluating visual-instruction alignment, and *Response Information Consistency (RIC)*, assessing response coherence relative to the instruction. Leveraging DIC, we propose Data Intrinsic Consistency Selection (DICS), an adaptive selection method balancing intra-sample consistency with distributional diversity across varying data budgets. Extensive experiments demonstrate that DICS consistently outperforms SoTA baselines across diverse dataset scales and model architectures, validating its robustness and scalability.

Limitations and Future Work. A primary limitation is that our consistency evaluation currently focuses on image-text data, excluding modalities like video or audio; we plan to extend this framework to broader multimodal contexts in future work. We identify two promising directions: (1) developing an adaptive strategy to automatically determine optimal subset size and composition, and (2) shifting from discarding low-DIC samples to actively repairing them to enhance model performance.

Ethical Statement. In this work, we utilize publicly available visual instruction tuning datasets for visual instruction selection to facilitate reproducibility. Concurrently, following the data curation protocols outlined in InternVL, we have compiled an additional set of data, which will be publicly released in the near future (see Appendix B.1). However, certain samples within these datasets contain erroneous answers regarding the images that lack a clear alignment with the provided instructions and responses. Fine-tuning Large Vision-Language Models (LVLMs) on such flawed data can cause the models to generate incorrect visual interpretations or suffer from hallucinations. To address this issue, our proposed method aims to effectively choose high-quality samples, thereby improving the efficacy of instruction tuning. Therefore, our work does not introduce any unforeseen ethical concerns.

References

Jinhe Bi, Yifan Wang, Danqi Yan, Wenke Huang, Zengjie Jin, Xiaowen Ma, Artur Hecker, Mang Ye, Xun Xiao, Hinrich Schuetze, and 1 others. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Zhe Chen, Weiyun Wu, Yue Wang, Wenhai Liu, Xinghang Zhang, Huipeng Sun, Zehui Chen, Jianbo Yang, Pan Zhang, Jiaqi Li, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems (Neurlps)*, 36:49250–49267.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Neurlps)*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3608–3617.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)*, pages 7514–7528.

Yuyang Hong, Qi Yang, Tao Zhang, Zili Wang, Zhaojin Fu, Kun Ding, Bin Fan, and Shiming Xiang. 2025. Taming modality entanglement in continual audio-visual segmentation. *arXiv preprint arXiv:2510.17234*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. In *2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5060–5080. Association for Computational Linguistics.

Seulbi Lee and Sangheum Hwang. 2026. Selective training for large vision language models via visual information gain. *arXiv preprint arXiv:2602.17186*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning (ICML)*, pages 19730–19742. PMLR.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. 2023b. Reflection-tuning: Data recycling improves llm instruction-tuning. *arXiv preprint arXiv:2310.11716*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (NAACL)*, pages 7602–7635.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems (Neurlps)*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024a. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision (ECCV)*, pages 216–233. Springer.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024b. Less is more: Data value estimation for visual instruction tuning. *CoRR*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems (Neurlps)*, 35:2507–2521.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2021. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems (Neurlps)*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PmLR.

Thiziri Nait Saada, Louis Bethune, Michal Klein, David Grangier, Marco Cuturi, and Pierre Ablin. 2025. The data-quality illusion: Rethinking classifier-based quality filtering for llm pretraining. *arXiv preprint arXiv:2510.00866*.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656.

Zinan Tang, Xin Gao, Qizhi Pei, Zhuoshi Pan, Mengzhang Cai, Jiang Wu, Conghui He, and Lijun Wu. 2025. Middo: Model-informed dynamic data optimization for enhanced llm fine-tuning via closed-loop learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6882–6902.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2024. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 54104–54132.

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 (ACL)*, pages 15271–15342.

Xu Yang, Chen Liu, and Ying Wei. 2025. Data selection matters: Towards robust instruction tuning of large multimodal models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (Neurlps)*.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2025. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 155–165.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *arXiv preprint arXiv:2402.04833*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Implementation Details

A.1 Token Weighting Strategy

To enhance semantic sensitivity in both VIC (Sec. 3.1) and RIC (Sec. 3.2), we employ a heuristic-based token weighting strategy. Specifically, we assign a full weight of 1.0 to content words to prioritize semantics, and a reduced weight of 0.1 to functional words to prevent frequent but grammatically necessary tokens from dominating the calculation.

We utilize the spaCy library (Honnibal et al., 2020) to automatically identify the Part-of-Speech (POS) tag for each token and assign these weights accordingly. Formally, let $\text{POS}(t_i)$ denote the part-of-speech tag of token t_i. We define the weight w_i as:

$$w_i = \begin{cases} 1.0, & \text{if POS}(t_i) \in \mathcal{C}, \\ 0.1, & \text{if POS}(t_i) \in \mathcal{F}, \\ 1.0, & \text{otherwise.} \end{cases} \tag{7}$$

where $\mathcal{C}$ and $\mathcal{F}$ denote content words and functional words, respectively.

Content Words ($\mathcal{C}$). These are semantically meaningful tokens that carry the core information of a sentence. To prioritize their impact, we assign full weight ($w_i = 1.0$) to the following POS categories:

- **Nouns** (NN, NNS, NNP, NNPS): such as "dog", "image", and "building";
- **Verbs** (VB, VBD, VBG, VBN, VBP, VBZ): such as "describe", "show", and "contains";
- **Adjectives** (JJ, JJR, JJS): such as "red", "large", and "beautiful";
- **Adverbs** (RB, RBR, RBS): such as "clearly", "very", and "quickly";
- **Numbers** (CD): such as "three" and "2024".

Functional Words ($\mathcal{F}$). These tokens primarily serve grammatical purposes and contribute less to the semantic content. To prevent them from dominating the metric, we assign reduced weight ($w_i = 0.1$) to:

- **Determiners** (DT): such as "the", "a", "an", and "this";
- **Prepositions** (IN): such as "in", "on", "at", and "of";
- **Conjunctions** (CC): such as "and", "but", and "or";
- **Pronouns** (PRP, PRP$, WP, WP$): such as "it", "they", and "which";
- **Auxiliary Verbs** (MD): such as "can", "will", and "should";
- **Punctuation** : such as ".", ",", ":", etc;
- **Particles** (RP, TO): such as "to" and "up".

This rigorous weighting scheme ensures that our proposed metrics effectively capture semantically significant elements while filtering out noise from informationally sparse components.

A.2 Implementation Details of VIC

Image Blocking. To compute the loss $\mathcal{L}_w(y \mid x)$ without visual information, we need to remove the influence of the image while preserving the input sequence structure. Our default approach is to **mask out the image tokens** from the input sequence, preventing any visual information from being attended to during response generation. This ensures a clean ablation of visual information without introducing distributional artifacts. We also explored an alternative approach: **replacing the original image with an unrelated image** randomly sampled from the dataset. This substitution strategy disrupts the semantic alignment between the image and the instruction-response pair while maintaining the natural distribution of visual features. Empirically, both approaches yield comparable results, suggesting that the VIC metric is robust to the specific implementation of visual information removal.

Text-only Samples. For samples without images (i.e., pure text instruction data), the VIC score is 1.0 according to Eq. 3, indicating neutral visual information gain.

Multi-image Samples. When multiple images are present, all images are processed together as the visual context I. The VIC metric measures the collective information gain from all images.

A.3 Implementation Details of RIC

As discussed in Sec. 3.2, the inverse formulation is essential to explicitly penalize generic, low-information responses (e.g., "This is an image."), which might achieve trivially low loss in standard forward prediction despite lacking specific relevance to the instruction. To compute the RIC score,

we guide the model to generate the original instruction x and measure the cross-entropy loss on its output tokens under the following two prompt templates:

- **Without Response** (for computing $\mathcal{L}_w(x \mid I)$): "`Based on the image, what is the likely question?`"
- **With Response** (for computing $\mathcal{L}_w(x \mid y, I)$): "`Based on the image and answer, what is the likely question? Answer: {y}`"

This design ensures that only responses providing instruction-specific semantic cues attain high RIC scores.

A.4 Implementation Details of DICS

In this section, we provide the detailed algorithmic procedures for our proposed Data Intrinsic Consistency Selection (DICS) method, which synergistically combines quality-based ranking with diversity-aware sampling.

Data Intrinsic Consistency Selection (DICS). As outlined in Section 4, DICS aims to select a high-quality and diverse subset from the original dataset based on DIC scores. The complete procedure is detailed in Algorithm 1. First, we compute the DIC score for each sample, defined as the minimum of its VIC and RIC scores, and sort the dataset in descending order. Then, we apply an adaptive selection strategy based on the target selection ratio p. For a low ratio ($p < 0.5$), we construct a candidate pool of size $2p \cdot |\mathcal{D}|$ containing the top-ranked samples, and perform diversity-aware sampling to select the final subset. For a high ratio ($p > 0.5$), we directly retain the top $(2p-1) \cdot |\mathcal{D}|$ samples as a core set, and apply diversity-aware sampling on the remaining samples to rescue the most diverse ones to meet the budget. At the boundary ($p = 0.5$), we simply select the top 50% samples.

Farthest Point Sampling (FPS). To ensure diversity within the selected subsets, DICS utilizes Farthest Point Sampling (FPS) as a core subroutine, detailed in Algorithm 2. FPS is a greedy algorithm that iteratively selects the sample most distant from the currently selected set, thereby maximizing the coverage of the embedding space. By applying FPS on dynamically narrowed candidate pools (as dictated by Algorithm 1) rather than the full dataset, our approach effectively balances data quality (prioritizing high DIC scores) and semantic diversity.

Algorithm 1 Data Intrinsic Consistency Selection (DICS).

Require: Dataset $\mathcal{D}$, selection ratio $p \in (0, 1]$
Ensure: Selected subset $\mathcal{D}^*$ with $|\mathcal{D}^*| = p \cdot |\mathcal{D}|$
Compute $s(d) = \min(\text{VIC}(d), \text{RIC}(d))$ for each $d \in \mathcal{D}$
Sort $\mathcal{D}$ by $s(d)$ descending $\rightarrow \mathcal{D}_{\text{sorted}}$
$n \leftarrow |\mathcal{D}|, \quad k \leftarrow p \cdot n$
if $p = 0.5$ **then**
  $\mathcal{D}^* \leftarrow$ top-k from $\mathcal{D}_{\text{sorted}}$
else if $p < 0.5$ **then**
  $\tilde{\mathcal{D}} \leftarrow$ top-$2k$ from $\mathcal{D}_{\text{sorted}}$
  $\mathcal{D}^* \leftarrow \text{FPS}(\tilde{\mathcal{D}}, k)$
else
  $\mathcal{D}_{\text{core}} \leftarrow$ top-$(2k - n)$ from $\mathcal{D}_{\text{sorted}}$
  $\tilde{\mathcal{D}} \leftarrow \mathcal{D}_{\text{sorted}} \setminus \mathcal{D}_{\text{core}}$
  $k_r \leftarrow k - |\mathcal{D}_{\text{core}}|$
  $\mathcal{D}_{\text{rescue}} \leftarrow \text{FPS}(\tilde{\mathcal{D}}, k_r)$
  $\mathcal{D}^* \leftarrow \mathcal{D}_{\text{core}} \cup \mathcal{D}_{\text{rescue}}$
end if
return $\mathcal{D}^*$

Furthermore, this piecewise sampling strategy significantly reduces the computational overhead compared to running FPS on the entire dataset.

Algorithm 2 Farthest Point Sampling (FPS)

Require: Candidate set $\tilde{\mathcal{D}}$ with embeddings $\{\mathbf{e}_i\}_{i=1}^{M}$, target size k
Ensure: Selected indices $\mathcal{S}$ with $|\mathcal{S}| = k$
Initialize $\mathcal{S} \leftarrow \{\arg\max_i \|\mathbf{e}_i\|\}$
Initialize distance array $d_i \leftarrow \|\mathbf{e}_i - \mathbf{e}_{\mathcal{S}[0]}\|_2$ for all i
for $j = 2$ to k **do**
  $i^* \leftarrow \arg\max_{i \notin \mathcal{S}} d_i$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{i^*\}$
  for $i \notin \mathcal{S}$ **do**
    $d_i \leftarrow \min(d_i, \|\mathbf{e}_i - \mathbf{e}_{i^*}\|_2)$
  end for
end for
return $\mathcal{S}$

Complexity Analysis. Let $M = |\tilde{\mathcal{D}}|$ denote the candidate pool size, k denote the target selection size, and d denote the embedding dimension. The time complexity of FPS is $O(k \cdot M \cdot d)$: each of the k iterations requires computing distances to all M candidates, with each distance computation taking $O(d)$ time. The space complexity is $O(M \cdot d)$ for storing embeddings plus $O(M)$ for the distance array.

In practice, we leverage GPU-accelerated batched distance computation to significantly reduce the wall-clock time. For a dataset of one million samples with 4096-dimensional embeddings, selecting 100K samples typically completes in minutes on a single A100 GPU.

B Experiment Details

B.1 Datasets

To strictly validate the effectiveness, robustness, and scalability of our proposed DICS method, we utilize three datasets spanning different data scales, distributions, and task types. All data collection complies with license requirements.

LLaVA-1.5-665K (Liu et al., 2023) is a comprehensive large-scale visual instruction tuning dataset. It aggregates 665,298 samples from diverse sources to enhance the model's multi-modal capabilities. Specifically, it consists of: (1) 158K GPT-generated multimodal instruction-following data from LLaVA-1.0-Instruct, (2) academic VQA datasets including VQAv2, GQA, and OK-VQA for visual reasoning, (3) OCR-related data from OCRVQA and TextCaps to bolster text-recognition abilities, and (4) 40K pure-language conversations from ShareGPT to maintain linguistic proficiency.

Vision-FLAN-186K (Xu et al., 2024) is a large-scale, multi-task visual instruction dataset specifically designed to address the limited task diversity in existing benchmarks. Unlike LLaVA-1.5-665K, which predominantly relies on conversational data generated by LLMs, Vision-FLAN is constructed by integrating 191 diverse vision-language tasks from various academic sources. It comprises 186,060 instances that cover a wide range of capabilities, including fine-grained recognition, spatial reasoning, and complex scene understanding.

DICS-6M is a large-scale visual instruction tuning corpus used to validate whether DICS remains effective in million-level data selection scenarios. Following the data collection protocol of InternVL (Chen et al., 2024c; Zhu et al., 2025), we collect 6M public high-quality vision-language SFT samples from widely used open-source datasets. As detailed in Table 7, the corpus covers a broad range of capabilities, including general vision, document understanding, mathematical reasoning, OCR, chart understanding, GUI, etc. The category distribution in Figure 6 further shows that the corpus is not dominated by a single narrow source, which supports a rigorous and reliable evaluation of large-scale data selection.

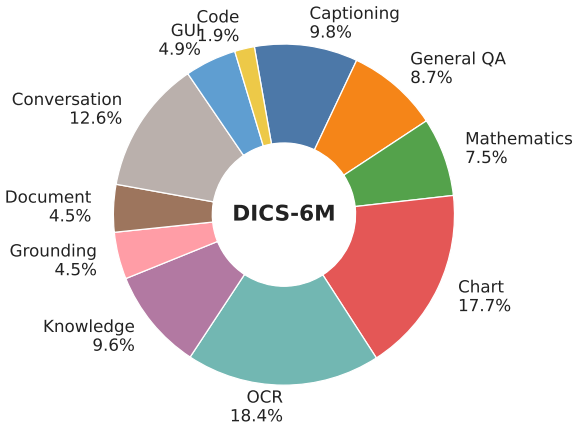

Figure 6: **Task distribution of the DICS-6M instruction dataset.** The collected corpus encompasses diverse tasks with a balanced distribution, facilitating a robust evaluation of large-scale data selection.

B.2 Benchmarks

We conduct a comprehensive evaluation to assess the multi-dimensional capabilities of the trained models. Following the standard evaluation protocols of LLaVA (Liu et al., 2023), we extend the benchmark suite to include a broader range of domains. To the best of our knowledge, our evaluation covers the most extensive set of benchmarks among current multi-modal data selection studies, categorized into five distinct domains:

- **General Capability:** We use **MME** (Fu et al., 2025) and **MMBench** (Liu et al., 2024a) (both English and Chinese versions). MME evaluates perception and cognition across 14 subtasks, while MMBench employs a circular evaluation strategy to robustly test model capabilities. We normalize the scores on MME for computing average performance.

- **Knowledge & Reasoning:** We employ **MMMU** (Yue et al., 2024) for college-level multidisciplinary reasoning requiring domain-specific knowledge. **SQA-I (ScienceQA-Image)** (Lu et al., 2022) assesses scientific reasoning with multi-modal contexts, and **VizWiz** (Gurari et al., 2018) tests the model's ability to answer visual questions originating from blind users, reflecting real-world applicability.

- **Document Understanding:** Recognizing the importance of fine-grained text perception, we

Table 7: **Composition of the DICS-6M instruction dataset.**

Task	Datasets
Captioning	TextCap, LVIS-Instruct4V, ShareGPT4V, ShareGPT4o, InternVL, COCO, NewYorker, DenseFusion, OCR, chinesememe, GPT-4o, Memotion, Edraw, Clothes Caption, AIG Share
General QA	GQA, OKVQA, VQAv2, Visual7W, VSR, Objects365, IconQA, VQAonBD, Hateful Memes, FSC147, K12 Printing, CLEVR, LNQA, RobuT, COCO, Vision-FLAN, GPT-4o, Image Textualization, RAVEN, COCO-QA, VizWiz, Douban, Vision Oriented, CoSyn, LLRV, IDK, Ctrip, SketchyVQA, OODVQA, FinQA, VQA-AS, YesBut, SPARK, SVRD, Indoor QA, AIG Share, study_com, gpt4v
Mathematics	MAVIS, MapQA, GeoQA, Geometry3K, UniGeo, GEOS, CLEVR-Math, GeomVerse, LaTeX Formulas, SynthFormulaNet, MathWriting, TallyQA, Geo170K, CoSyn, CLEVR, Gaokao, InterGPS, PicMath, Geo3K
Chart	ChartQA, PlotQA, FigureQA, LRV-Instruction, ArxivQA, MMC-Inst, TabMWP, DVQA, UniChart, SimChart9K, Chart2Text, FinTabNet, SciTSR, TinyChart, SynthChartNet, ECD, MMTab, Datik, CoSyn, SBT Chart, Wired Table, VisText, HiTab, Oroikon, TAT-QA, Charts2500
OCR	OCRVQA, InfoVQA, TextVQA, ArT, HME100K, COCO-Text, CTW, LSVT, RCTW-17, VCR, EST-VQA, ST-VQA, EATEN, LLaVAR, CASIA, Chinese-OCR, IAM, NAF, POIE, ReCTs, MTWI, TextOCR, SROIE, Synthetic ArXiv OCR, Synthetic Infographic2Markdown, Synthetic OCR, WIT, SynthDog, TAL-OCR, K12 Printing, olmOCR, OCR, Captcha, LaTeX-OCR, CyrillicHandwriting, LaTeX Handwritten, ICDAR 2019, HW-SQuAD, Rendered Text, Chrome Writing, AIG Share, LaTeX QA, Invoices & Receipts, MTVQA, imgur5k, ORAND-CAR, IIIT5K, Handwriting Forms, Thai OCR, MapText
Knowledge	KVQA, A-OKVQA, ViQuAE, iNaturalist2018, MovieNet, ART500K, KonIQ-10K, TQA, ChemVLM, ScienceQA, AI2D, VQA-RAD, Wikipedia, Google Landmarks, Chinese Culture, Face Emotion, Diagram, CoSyn, Gaokao, VisualMRC, BlockDiagram
Grounding	COCO-ReM, V3Det, All-Seeing-V2, TolokaVQA, GPT4Gen-RD-BoxCoT, VisualWebInstruct, Downstream Grounding, Localized Narratives, OpenApp, AIG Share, SpatialSense
Document	Sujet-Finance-QA-Vision, UReader, allenai_pixmo_docs, ArXiv Figures, CoSyn, InfographicVQA, Bentham, PDF-VQA, OmniDocBench, Layout Extract, Schedule Extract, FUNSD
Conversation	ALLaVA, SVIT, Cambrian, LAION-GPT4V, WildVision, Viet-ShareGPT4o, llava_instruct, vflan, MMEvol, llava_cot_100k, llava_wild, RLAIF-V, Visual Chat, GPT-4o, AIG Share
GUI	Screen2Words, WebSight, Widget-Caption, RICOSCA, SeeClick, ScreenQA, AMEX, Android UI, UIBert, WaveUI, RootsAutomation, AlfredPLPL, Phone Action, Web Collected, AlfWorldGPT, Home Screen, Taobao App, Airplane App, WeChat App
Code	SynthCodeNet, Datikz, Drawing2HTML

include **AI2D** (Kembhavi et al., 2016) for diagram understanding, **DocVQA** (Mathew et al., 2020) for document comprehension, and **InfoVQA** (Mathew et al., 2021) for infographic reasoning.

- **Hallucination:** We focus on model reliability using **POPE** (Li et al., 2023c), which evaluates object existence hallucination via polling, and **HallusionBench** (Guan et al., 2024), which focuses on visual illusions and reasoning consistency.

- **Open-ended Dialogue:** We utilize **MM-Vet** (Yu et al., 2023) to evaluate integrated capabilities in solving complex, open-ended problems that require reasoning, recognition, and generation.

Evaluation Protocol. For all evaluations, we utilize VLMEvalKit (Duan et al., 2024), a standardized and reproducible evaluation framework. To ensure deterministic and fair comparisons, we set the generation temperature to 0.0 for all benchmarks.

For benchmarks requiring an LLM-based judge (e.g., MM-Vet and open-ended responses in MM-Bench variants), we employ Qwen3-32B as the judge model. This choice ensures a high correlation with human judgment while maintaining evaluation efficiency.

B.3 Implementation Details of Baselines

We compare our framework against a comprehensive set of data selection strategies, categorized into four groups: (1) **Stochastic Sampling**: *Random Selection* serves as the fundamental baseline, uniformly sampling a fixed proportion of data from the instruction set. (2) **Heuristic Filtering**: This category includes *Length-based filtering*, which prioritizes samples based on sequence length, and *CLIP-Score* (Radford et al., 2021), which selects data based on coarse-grained image-text alignment, where we utilize the clip-vit-large-patch14 model to retain samples with the highest image-question similarity. (3) **LLM Selectors**: We evaluate strategies relying on intrinsic language model properties, including *Perplexity* (Marion et al., 2023), which estimates sample probability to filter out abnormal or high-noise data (employing LLaVA-1.5-7B for mid-range selection following prior work), and *IFD* (Li et al., 2024b), which assesses the instruction following difficulty by measuring the loss differential between instruction-conditioned and unconditioned responses. (4) **MLLM Selectors**: We compare against advanced strategies specifically designed for multimodal instruction tuning, including *COINCIDE* (Lee et al., 2024), *DataTailor* (Yu et al., 2025) and *PRISM* (Bi et al., 2025). We strictly follow the official implementations and default settings for all advanced baselines.

C Experiments and Analysis

In this section, we provide a comprehensive analysis to complement the experimental results presented in Section 5. Specifically, we delineate performance across individual benchmarks to ensure full transparency. Furthermore, we present extended comparative experiments and ablation studies to corroborate the robustness and effectiveness of our proposed DICS framework.

C.1 More Results

Superiority across Data Scales. As shown in Figure 1, DICS consistently outperforms all methods across varying sampling ratios. In low-data ratios (<25%), it establishes a distinct lead, validating that our consistency metrics effectively identify high-value samples. Notably, DICS achieves peak performance at the 50% ratio (103.15%). While extending to high-data regimes (>75%) leads to generic performance saturation or slight decline across all methods, DICS maintains significantly greater stability and resilience. In contrast, baselines like PRISM (Bi et al., 2025) experience sharper degradation due to the inclusion of low quality samples. This confirms DICS's ability to maximize data efficiency while effectively filtering out inconsistent samples. Notably, p is not fixed but depends on the specific metric and dataset; other sample-level methods also exhibit such peaks (e.g., 75% for PRISM and DataTailor in Figure 1).

Distributions of DIC Scores. Figure 7 illustrates the distribution of VIC and RIC scores within the selected 25% subset of LLaVA-1.5-665K dataset. The samples are concentrated in the higher value ranges for both metrics, confirming that our method effectively prioritizes data with strong visual dependency and logical coherence. Simultaneously, the distribution exhibits a broad coverage across these high scores rather than collapsing into a narrow peak. This healthy variance indicates that samples of varying complexity are retained, effectively preventing homogeneity. This balance ensures that the final dataset achieves high intrinsic consistency without sacrificing the distributional diversity required for robust model generalization.

Statistical Analysis. To further validate the statistical significance and robustness of our proposed framework, we report the mean scores and standard deviations over 5 independent runs for the main experiments corresponding to Table 1 and Table 2.

Figure 7: **The distribution of DIC scores of 25% selected data of LLaVA-1.5-665K dataset.**

The detailed results are summarized in Table 9. As shown in the table, the variances across all methods are extremely minimal (all standard deviations ≤ 0.13). Specifically, the performance gaps between our DICS framework and the strongest baseline, PRISM (e.g., +1.79 in Table 1 and +0.63 in Table 2), are substantially larger than the standard deviations. This confirms that our improvements are statistically significant and clearly exceed random fluctuations. This overall stability thoroughly demonstrates our method's robustness and proves that the performance gains are intrinsic to our data selection strategy.

C.2 Robustness and Scalability

Based on the granular results presented in Table 10, we provide a deeper analysis of the effectiveness of DICS across different model architectures and scales.

Architecture-Agnostic DIC Metrics. A key question in data selection is whether the filtered subset is biased towards the selector model's specific preferences or if it captures intrinsic data quality. Observing the **Qwen2-VL-7B** target experiments, we find that data selected by the **LLaVA-1.5-7B** model achieves a relative performance of **101.25%**, slightly outperforming the subset selected by Qwen2-VL-7B itself (100.91%) and significantly surpassing the full dataset baseline. This indicates that DICS captures *universal* consistency between visual and textual modalities rather than model-specific patterns. The high transferability implies that we can utilize a single, well-established model (like LLaVA) to curate data for different architectures without re-computing scores for each new target model.

Scalability via Weak-to-Strong Generalization. The results on the **LLaVA-1.5-13B** target model offer compelling evidence for the scalability of our approach. Training the 13B model using data se-

Table 8: **Detailed performance comparison on Vision-FLAN dataset.** All methods are evaluated by selecting a 25% subset of the training data.

Method	General Capability			Knowledge & Reasoning			Document Understanding			Hallucination		Dialogue	
	MME	MMBench en	MMBench cn	MMMU	SQA-I	VizWiz	AI2D	DocVQA	InfoVQA	POPE	Hallusion Bench	MMVet	Rel. (%)
Full Dataset	53.93	53.79	48.07	35.78	64.45	25.99	51.20	20.61	16.82	83.85	32.57	35.28	100.00
Random Selection	**53.45**	47.29	42.96	35.11	63.91	25.34	44.85	21.73	18.43	80.85	**34.89**	34.40	96.34
Length (Zhao et al., 2024)	53.07	43.81	38.93	34.22	58.50	23.99	39.54	16.22	17.90	81.93	33.57	36.10	91.47
CLIP-Score (Hessel et al., 2021)	51.58	42.96	38.85	33.78	**65.94**	25.46	**52.04**	25.11	**23.10**	82.56	32.91	**37.06**	97.90
Perplexity (Marion et al., 2023)	39.59	29.02	22.52	30.22	57.91	26.72	35.10	26.13	22.57	79.54	33.47	33.35	83.50
IFD (Li et al., 2024b)	51.97	40.09	40.33	33.56	65.15	20.18	49.74	16.08	20.95	82.26	29.63	36.38	93.11
DataTailor (Yu et al., 2025)	52.98	42.41	38.54	31.67	65.10	25.83	46.99	25.29	20.65	81.91	34.60	36.38	96.18
PRISM (Bi et al., 2025)	52.85	**48.84**	**45.74**	**35.89**	65.15	25.73	50.55	22.14	18.14	83.04	34.04	32.29	98.48
DICS (Ours)	50.67	48.22	44.97	34.78	65.05	**26.93**	50.55	**26.63**	20.41	**84.44**	32.15	**37.06**	**99.91**

Table 9: **Mean scores and standard deviations over 5 independent runs.** The results correspond to the overall performance in Table 1 and Table 2.

Method	Score (Mean $\pm$ Std)	
	For Table 1	For Table 2
Full Dataset	46.29 ± 0.03	43.47 ± 0.06
Random Selection	44.64 ± 0.13	41.85 ± 0.07
DataTailor	44.86 ± 0.10	41.94 ± 0.07
PRISM	45.18 ± 0.05	42.90 ± 0.03
DICS (Ours)	$\mathbf{46.97 \pm 0.04}$	$\mathbf{43.53 \pm 0.05}$

lected by the much smaller **7B** selector yields a relative performance of **101.79%**, the highest improvement margin among all settings. This "Weak-to-Strong" generalization capability is practically significant. It suggests that DICS enables a computationally efficient pipeline where lightweight models act as "data filters" for training large-scale models. This significantly reduces the computational overhead of the data selection phase, which is traditionally a bottleneck when processing massive datasets. Consequently, this transferability suggests that well-established models (like LLaVA) have the potential to serve as efficient data curators for diverse architectures, reducing the computational redundancy of re-calculating scores for every new target model.

Architectural Generalizability of DICS. To strictly validate the quality of our selected subsets, we compare our models (fine-tuned from **Qwen2-VL-7B-Base**) against the official checkpoints. As shown in the table 10, the Qwen2-VL-7B-Base model exhibits limited instruction-following capabilities (Score: 51.05), serving as a lower-bound baseline. However, by fine-tuning on just 25% of the LLaVA-665K dataset selected by DICS, our model achieves a substantial performance leap to **71.35**. Remarkably, this result bridges the gap towards the official **Qwen2-VL-7B-Instruct** model (Score: 72.20) to within a 1% margin. This demonstrates that prioritizing data intrinsic consistency allows for highly efficient model training with minimal resource consumption.

C.3 Ablation Studies and Analysis

Effectiveness of DIC. We dissect the contribution of each DIC component to understand their roles in data selection in Table 11.

- **Necessity of Quality Filtering:** The "Diversity Only" baseline (selecting data solely based on feature embedding distance) yields the lowest performance (97.02%). This confirms that maximizing diversity without ensuring intrinsic quality introduces low-quality samples that degrade model capabilities.

- **Complementarity of VIC and RIC:** VIC serves as the primary driver for general capabilities, achieving a high relative score of 101.02%. In contrast, while RIC alone underperforms in general tasks (97.60%), it excels in hallucination benchmarks like **POPE** and **HallusionBench**. Consequently, VIC ensures the fidelity of visual and textual content, whereas RIC promotes structural and logical consistency, effectively reducing the incidence of hallucinations in complex multimodal tasks.

- **Fusion Strategies:** Simple fusion strategies like Sum or Product fail to outperform the single VIC metric. However, our proposed **min(VIC, RIC)** operator achieves the best overall performance (101.40%). The "min" operator acts as a strict gate, requiring samples to be high quality in *both* visual alignment and response logic, effectively filtering out partial outlier samples.

- **Impact of Token Weighting:** Removing the

Table 10: **Detailed generalization analysis across architectures and scales.** This table provides the granular results for the cross-architecture experiments for Table 3. *Model* indicates the backbone being trained, while *Selector* indicates the model used to calculate DIC scores for data filtering. All subset selection methods employ a 25% sampling ratio. The best results for each target model are highlighted in **bold**, and second-best are underlined.

Model	Selector	General Capability			Knowledge & Reasoning			Document Understanding			Hallucination		Dialogue	
		MME	MMBench en	MMBench cn	MMMU	SQA-I	VizWiz	AI2D	DocVQA	InfoVQA	POPE	Hallusion Bench	MMVet	Rel. (%)
LLaVA-1.5-7B	Full Dataset	59.92	60.22	51.93	35.89	**68.32**	31.79	47.91	23.82	22.30	84.49	27.80	41.01	100.00
	Random Selection	59.28	56.27	51.93	35.56	66.98	29.53	50.91	21.62	21.63	81.58	**29.46**	38.07	97.74
	DICS (LLaVA-1.5-7B)	**63.39**	**59.52**	**55.03**	**36.11**	66.14	**31.29**	**50.52**	**24.64**	**23.11**	**84.91**	26.79	41.74	**101.40**
	DICS (Qwen2-VL-7B)	62.27	58.98	54.57	36.00	65.54	31.22	48.87	24.24	23.17	83.78	25.19	**42.48**	100.16
Qwen2-VL-7B	Full Dataset	78.06	78.72	78.02	48.11	81.90	41.88	80.83	91.92	73.14	87.97	**47.18**	57.84	100.00
	Random Selection	**79.51**	78.72	76.70	46.67	81.51	41.87	81.28	92.72	74.27	86.37	46.64	**61.10**	100.21
	DICS (LLaVA-1.5-7B)	78.17	**79.10**	**79.72**	**52.00**	82.10	**42.28**	**81.28**	92.32	**74.50**	**88.51**	45.15	61.01	**101.25**
	DICS (Qwen2-VL-7B)	77.39	78.64	78.25	**52.00**	**83.09**	41.73	81.02	**92.37**	73.79	87.68	46.64	60.64	100.91
Ref: Qwen2-VL-7B-Base	-	67.94	75.77	73.99	50.33	81.71	1.94	79.34	4.75	0.78	78.98	42.42	54.68	72.45
Ref: Qwen2-VL-7B-Instruct	-	81.80	79.18	79.49	47.00	84.23	42.46	82.51	93.83	75.51	85.82	50.58	63.94	102.46
LLaVA-1.5-13B	Full Dataset	54.06	**62.93**	56.81	**37.33**	69.16	**32.83**	**58.13**	26.39	**26.16**	81.96	24.80	44.17	100.00
	Random Selection	62.44	60.68	53.25	36.56	**71.84**	30.97	55.99	24.21	25.19	85.16	23.44	**44.59**	99.93
	DICS (LLaVA-1.5-7B)	**63.62**	61.76	**58.05**	37.11	70.10	32.06	56.02	26.04	25.90	**85.93**	**24.92**	43.49	**101.79**

Table 11: **Ablation study on Data Intrinsic Consistency (DIC).** We analyze the impact of individual consistency metrics, different fusion strategies, and token weighting. All methods use the LLaVA-1.5-665K dataset with a 25% sampling ratio on the LLaVA-1.5-7B model. The best results are highlighted in **bold**.

Metric Variant	General Capability			Knowledge & Reasoning			Document Understanding			Hallucination		Dialogue	
	MME	MMBench en	MMBench cn	MMMU (val)	SQA-I	VizWiz	AI2D	DocVQA	InfoVQA	POPE	Hallusion Bench	MMVet	Rel. (%)
Baselines													
Full Dataset	59.92	**60.22**	51.93	35.89	**68.32**	**31.79**	47.83	23.82	22.30	84.49	27.80	41.01	100.00%
FPS (Diversity Only)	58.51	54.72	49.23	33.67	66.58	29.94	50.45	22.78	22.35	84.08	27.44	39.08	97.02%
Single Consistency Metric													
w/ VIC only	**63.87**	59.06	54.18	**36.44**	66.63	31.50	**50.97**	24.24	22.96	83.92	27.72	39.54	101.02%
w/ RIC only	59.63	53.17	48.92	35.44	66.09	30.22	49.68	22.38	22.29	**85.83**	**28.13**	40.28	97.60%
Metric Fusion Strategies													
Sum (VIC + RIC)	59.01	57.12	53.17	33.33	66.78	31.31	48.64	24.10	21.75	84.67	26.22	41.56	98.61%
Product (VIC $\times$ RIC)	58.12	56.66	53.95	35.89	66.68	31.02	48.06	24.25	22.85	84.70	27.63	39.17	98.84%
Weighted Token Strategy													
DIC w/o weighted token.	61.02	58.59	54.95	34.33	66.29	31.17	50.58	24.60	**23.38**	84.35	27.49	40.05	100.25%
DIC (ours)	63.39	59.52	**55.03**	36.11	66.14	31.29	50.52	**24.64**	23.11	84.91	26.79	**41.74**	**101.40%**

POS-based weighting leads to a performance drop (101.40% $\rightarrow$ 100.25%), validating that focusing on content-rich tokens (nouns, verbs, etc) provides a more accurate assessment of semantic consistency.

Impact of Selection Strategies. Once the high-quality candidate set is identified, the choice of selecting strategy for the second stage is critical for maintaining distribution coverage. We detail the results in Table 12. As shown in Table 12, all diversity-aware strategies (Hierarchical, and FPS) consistently outperform the full dataset baseline (all >100%). This robustness confirms that the core contribution comes from the DIC-based quality filtering stage, which successfully identifies a high-value data distribution regardless of the subsequent selection method.

Visualization of VIC and RIC. To provide a more intuitive understanding of our proposed metric, we present representative samples with high and low DIC scores from the LLaVA-1.5-665K dataset in Figure 8. Samples with high Visual Information Consistency (VIC) scores are often characterized by distinct visual elements, well-defined questions and informative answers, demonstrating a strong cross-modal alignment between vision and language. Conversely, instances with low VIC scores often contain ambiguous or incorrect answers, where the visual input provides minimal information gain for the given question. Regarding Response Information Consistency (RIC), high-scoring samples generally exhibit rich descriptive content, clear reasoning logic, and a strong semantic correlation between the question and the answer. In contrast, samples with low RIC scores tend to produce vague, generic responses that lack informative content.

Qualitative Comparison via Attention Map. To generate the text-to-image attention maps, we perform a complete forward pass of the LLaVA-1.5 model (which consists of 32 layers in total) and

Table 12: **Ablation study on selection strategies.** We compare different diversity-aware selection methods applied to the high-quality subset filtered by DIC. All methods use the LLaVA-1.5-665K dataset with a 25% sampling ratio on the LLaVA-1.5-7B model. The best results are highlighted in **bold**.

Sampling Strategy	General Capability			Knowledge & Reasoning			Document Understanding			Hallucination		Dialogue	
	MME	MMBench en	MMBench cn	MMMU	SQA-I	VizWiz	AI2D	DocVQA	InfoVQA	POPE	Hallusion Bench	MMVet	Rel. (%)
Full Dataset	59.92	**60.22**	51.93	35.89	**68.32**	31.79	47.83	23.82	22.30	84.49	27.80	41.01	100.00%
K-Means (K=5,000)	61.58	58.44	54.72	35.89	65.94	**32.02**	**50.87**	24.44	**23.54**	84.59	25.62	40.37	100.47%
Hierarchical	**64.09**	57.20	53.79	**36.44**	67.48	31.91	49.90	**24.77**	23.15	83.97	**29.22**	41.19	101.39%
FPS (Ours)	63.39	59.52	**55.03**	36.11	66.14	31.29	50.52	24.64	23.11	**84.91**	26.79	**41.74**	**101.40%**

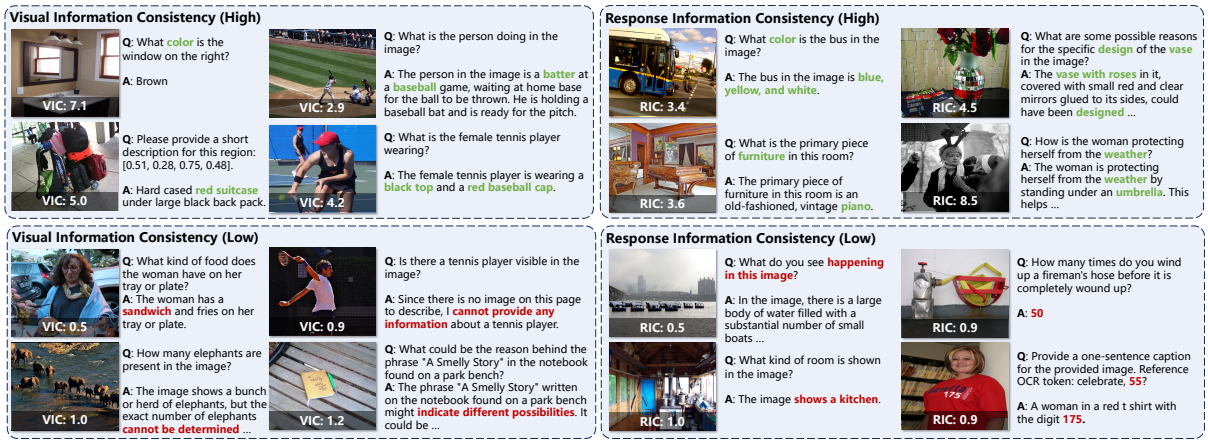

Figure 8: **More visualization of DIC scores.**

extract the attention weights from a specific layer of the language model (specifically, layer 21). We empirically select this intermediate-deep layer because it offers an optimal balance between spatial precision and semantic relevance. In previous studies and our observations, deeper layers tend to capture more abstract semantic information but lack fine-grained spatial localization, whereas shallower layers produce more diffuse and noisy attention patterns. We aggregate the attention weights from the generated response tokens to the visual patch tokens, yielding a 576-dimensional text-to-image attention distribution. This distribution is then reshaped into a 24×24 spatial grid, smoothed, and upsampled to the original image resolution using bilinear interpolation. Finally, we apply contrast enhancement and the JET colormap to generate the heatmap, which is overlaid onto the original image. The baseline for comparison is the model trained on a 15% randomly selected subset.

As illustrated in Figure 9, the model trained with our DICS framework demonstrates highly concentrated and accurate attention on the core target objects. For instance, when answering questions about specific visual details, the DICS model precisely localizes the relevant entities. In contrast, the random selection baseline often distributes its attention diffusely across the image or focuses on irrelevant background regions. This qualitative evidence further substantiates that DICS effectively encourages the model to rely on critical visual evidence rather than language priors.

Table 13: **Computational cost analysis on LLaVA-665K (N=665K) using 8×A100 GPUs.** The pipeline is dominated by inference-time costs ($O(N\mathcal{F})$), while the selection overhead is minimal.

Stage	Complexity	Time	Type
Warmup Training	$O(EN\mathcal{F})$	<1h	Gradient/Fwd
DIC & Embedding	$O(N\mathcal{F})$	∼10h	Inference Only
Data Selection	$O(N \log N + MND)$	<1h	Vector Ops

C.4 Computational Efficiency

We analyze the computational efficiency of our data selection pipeline. Let $\mathcal{F}$ denote the complexity of a single forward pass. Given a dataset of N samples, a target subset size M, and embedding dimension D, the pipeline's computational cost scales linearly with N. As summarized in Table 13, this linear scaling ensures high scalability. The detailed breakdown is as follows:

- *Warmup Training:* We perform parameter-efficient LoRA fine-tuning for E epochs.

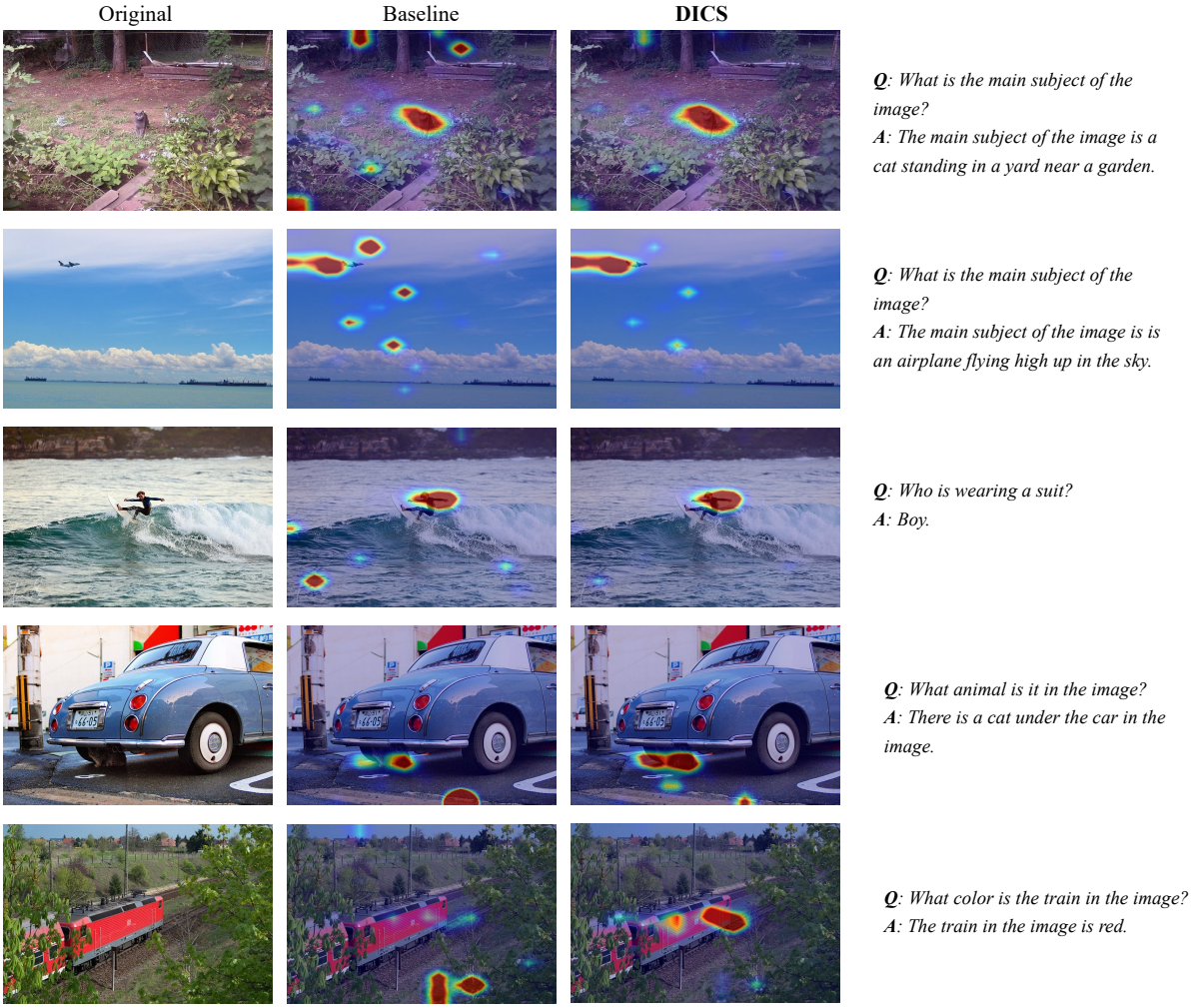

Figure 9: **Analysis of attention.** Compared to the random selection baseline, DICS encourages the model to pay more attention to the core visual details, rather than irrelevant background information.

Since LoRA updates only a fraction of parameters and the number of epochs is small, the complexity $O(EN\mathcal{F})$ remains computationally lightweight.

- *Metric & Embedding Computation:* To maximize efficiency, we integrate embedding extraction with metric computation. The visual and text embeddings are retrieved from the hidden states during the same forward passes used for VIC/RIC scores. This concurrent execution eliminates redundant model inference, maintaining a linear complexity of $O(N\mathcal{F})$ with negligible marginal cost for embedding storage.

- *Data Selection:* This stage involves score-based filtering ($O(N \log N)$) and Farthest Point Sampling ($O(MND)$). As $M \ll N$ and these operations involve only lightweight vector operations (e.g., matrix multiplication) without model invocation, the overhead here is negligible, allowing for rapid execution even on standard CPUs.