ICONS: Influence Consensus for Vision-Language Data Selection

Xindi Wu[1] Mengzhou Xia[1] Rulin Shao[2] Zhiwei Deng[3]
Pang Wei Koh[2,4] Olga Russakovsky[1]
[1]Princeton University [2]University of Washington [3]Google DeepMind [4]Allen Institute for AI
https://princetonvisualai.github.io/icons/

Abstract

Visual Instruction Tuning typically requires a large amount of vision-language training data. This data often containing redundant information that increases computational costs without proportional performance gains. In this work, we introduce ICONS*, a gradient-driven Influence CONsensus approach for vision-language data Selection that selects a compact training dataset for efficient multi-task training. The key element of our approach is cross-task influence consensus, which uses majority voting across task-specific influence matrices to identify samples that are consistently valuable across multiple tasks, allowing us to effectively prioritize data that optimizes for overall performance. Experiments show that models trained on our selected data (20% of* LLaVA-665K*) achieve 98.6% of the relative performance obtained using the full dataset. Additionally, we release this subset,* LLaVA-ICONS-133K*, a compact yet highly informative subset of* LLaVA-665K *visual instruction tuning data, preserving high impact training data for efficient vision-language model development.*

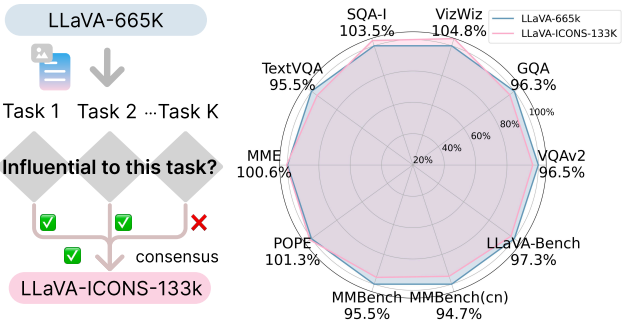


Figure 1. **Influence Consensus for Vision-Language Data Selection**. (*Left*) Given a large scale visual instruction tuning dataset (LLaVA-665K [23]), our method computes an influence score of each training example on each target task, aiming to select examples with high influence across multiple tasks. Our selected compact 20% subset (LLaVA-ICONS-133K) contains data points that achieve influence consensus across tasks, as indicated by the green check marks showing high influence ranking on corresponding tasks. (*Right*) The radar plot shows our subset achieves comparable performance across benchmarks, where percentage indicates relative performance to full dataset training.

1. Introduction

Visual instruction tuning has become a critical stage in developing Multimodal LLMs [22, 23], enabling these models to effectively interpret and respond to language instructions based on visual content. Current approaches typically rely on large instruction tuning datasets (e.g., LLaVA 1.5 [22] and Cambrian [39], which use 665K and 7M instruction tuning data points, respectively). These large datasets, while effective, pose practical challenges for research iteration and model development: extended development cycles due to lengthy training times [3, 15], increased storage and memory demands for managing large-scale datasets [8, 35], and substantial computational resources required for training [38, 40]. This motivates a fundamental question:

Can we identify a compact subset of training data that preserves model capabilities while enabling faster experimentation?

Prior work has explored various approaches to data selection, including gradient-based methods [6, 42], influence functions [17, 43], and diversity sampling [4, 44]. These methods typically either focus on optimizing for single tasks in isolation, or maximizing source data diversity without considering target task distributions. However, selecting data for visual instruction tuning requires careful consideration of performance across diverse downstream tasks. Simply selecting the most influential samples for an individual task risks creating a dataset that performs well on that specific benchmark but fails to develop broader capabilities. Rather than focusing on task-specific performance, we aim to identify universally valuable training samples by aggregating gradient-based influence measures across diverse tasks using a voting mechanism, effectively selecting data that contributes to broad model capabilities.

We introduce ICONS (Influence CONsensus vision-

language data Selection), a method that builds upon the gradient-based influence estimation approach from LESS [42]. Given access to validation data for each target task, our method: (1) computes first-order gradient influence scores to measure how each training sample impacts task-specific performance, and (2) uses influence consensus through majority voting to identify training samples that show consistent positive value across multiple tasks. This consensus-based mechanism identifies universally valuable training examples: while some samples might be highly influential for individual tasks, we prioritize those that demonstrate broad utility across the task spectrum. While the computational cost of influence estimation is expensive, this front-loaded, one-time investment yields a standardized, compact dataset that can significantly accelerate iteration and development of both multimodal data and models.

Using ICONS, we create LLaVA-ICONS-133K, an automatically curated 20% subset of the LLaVA dataset [23]. This compact dataset maintains 98.6% of the original performance across multiple vision-language tasks, providing a 2.8 percentage point improvement over randomly selecting the same-sized subset (95.8%) and eliminating two-thirds of the performance drop from shrinking the training data. Moreover, our ICONS outperforms all baselines across different selection ratios, and remarkably achieves above-full-dataset performance, surpassing the original dataset at a 60% selection ratio. Importantly, LLaVA-ICONS-133K shows strong transferability, maintaining 95.5-113.9% relative performance across diverse unseen tasks such as chart understanding and infographic comprehension, suggesting that ICONS identifies fundamentally valuable training examples rather than task-specific patterns. We summarize our key contributions:

1. We propose ICONS, a simple yet effective approach for multi-task vision-language data selection that identifies broadly valuable training samples through majority voting across task-specific gradient influence scores.
2. Our voting-based aggregation consistently outperforms all baselines (§3.2) and, most notably, exceeds 102% of the full dataset performance at 60% selection ratio (§3.3). We further analyze different influence aggregation approaches and show the effectiveness of our consensus-aware selection over direct-merge alternatives (§3.5).
3. We release LLaVA-ICONS-133K, a compact 20% subset of LLaVA-665K, achieving near-full performance (98.6%), transferring well to unseen tasks (§3.4), and serving as a standardized training set for resource-efficient development of multimodal models.

[Mengzhou: I think we can talk about the baselines a bit more in the introduction, are they all categorized as task-specific optimization? while other methods uses PPL/other methods as the specialist score without having a generalist step. I think it's worth making all these differences clear in the introduction.] [Xindi: fix1]

2. Influence Consensus for Vision-Language Data Selection

We propose a consensus-driven, gradient-based data selection framework for visual instruction tuning datasets. We formalize the problem setup in §2.1 and establish gradient-based influence estimation preliminaries in §2.2. Our two-stage data selection framework consists of: first, the *specialist* stage (§2.3), which computes task-specific influence scores, followed by the *generalist* stage (§2.4), which builds cross-task consensus through voting-based aggregation.

2.1. Problem Formulation

Given a large-scale visual instruction tuning dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{I}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ containing N samples, where each data point $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{I}_i, \boldsymbol{y}_i)$ includes natural language instruction $\boldsymbol{x}_i$ and an image $\boldsymbol{I}_i$, with corresponding target response $\boldsymbol{y}_i$[1], and given access to validation data $\mathcal{V}_k$ for each downstream task $T_k \in \mathcal{T} = \{T_1, ..., T_K\}$, our goal is to select a compact subset $\mathcal{S} \subset \mathcal{D}$ of size $M \ll N$ that maximizes model performance across multiple downstream tasks:

$$\mathcal{S}^* = \underset{\mathcal{S}\subset\mathcal{D},|\mathcal{S}|=m}{\arg\max} \sum_{k=1}^{K} \text{Rel}(f_{\mathcal{S}}, T_k), \tag{1}$$

$$\text{Rel}(f_{\mathcal{S}}, T_k) = \frac{\text{Score}(f_{\mathcal{S}}, T_k)}{\text{Score}(f_{\mathcal{D}}, T_k)}, \tag{2}$$

where $f_{\mathcal{S}}$ and $f_{\mathcal{D}}$ denote models trained on subset $\mathcal{S}$ and full dataset $\mathcal{D}$, respectively. $\text{Score}(f, T_k)$ is the task-specific evaluation score achieved by model f on task T_k. We define the average relative performance across all tasks as **Rel.** = $\sum_{k=1}^{K} \text{Rel}(f_{\mathcal{S}}, T_k)/K$. Rel. quantifies the subset-trained model's performance relative to that of the model trained on the entire dataset, with values close to 1 indicating that the subset maintains the performance of full training [18]. Our objective is to find a subset where Rel. $\approx$ 1, i.e., the model trained on the selected subset achieves performance comparable to using the full dataset across all tasks.

2.2. Preliminaries

Building on our problem formulation in §2.1, we formalize how to estimate the influence of training samples on downstream task performance. Specifically, since our goal is to maximize $\text{Rel}(f_{\mathcal{S}}, T_k)$ across tasks as defined in Eqn. 2, we need an efficient way to estimate how each training sample

[1]The framework supports multi-turn conversational data, yet we formalize the problem setup for single-turn instruction-tuning for clarity and simplicity.

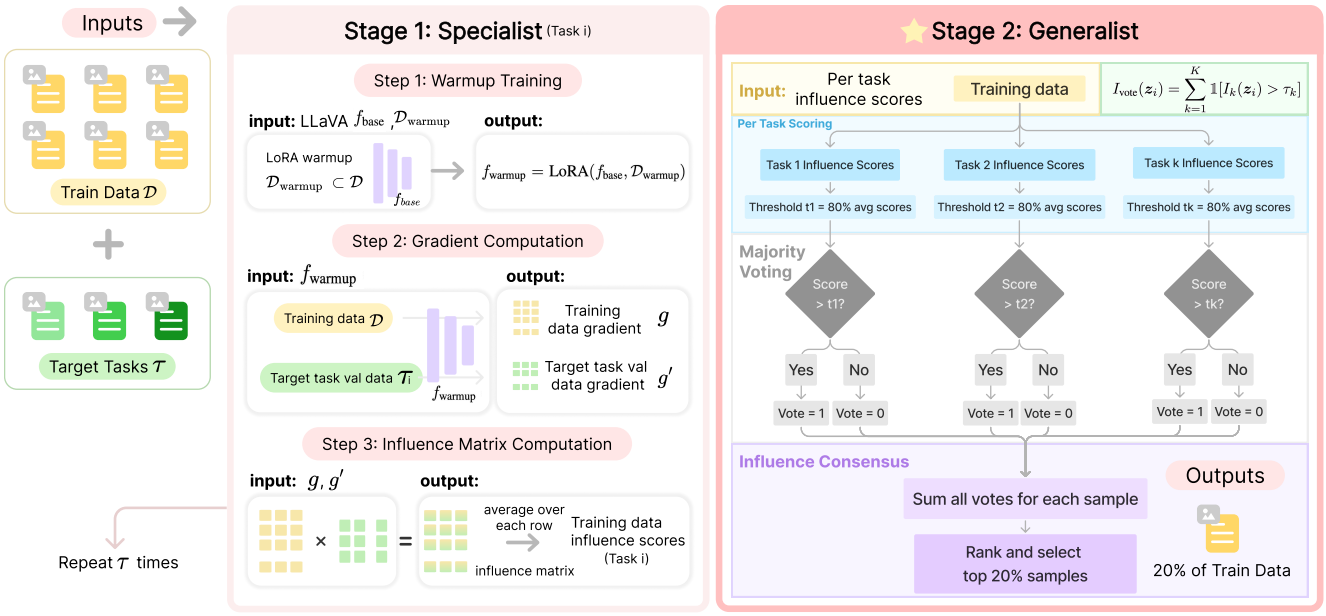


Figure 2. **ICONS.** The Specialist stage (*left*) processes each task individually through three steps: (1) warmup training on a small subset of data, (2) gradient computation for both training and target task validation data, and (3) influence matrix computation to generate per-task influence scores. This process is repeated for each target task. The Generalist stage (*right*) performs **Influence Consensus** to aggregate information across tasks, where samples scoring above the 80^{th} percentile threshold for each task receive a vote. The final selection is made by summing votes across tasks and selecting the top 20% most influential samples, creating a compact yet highly effective training dataset that performs well across all tasks.

contributes to the Score$(f_{\mathcal{S}}, T_k)$ term in the numerator. Following [33, 42], we estimate how individual training samples $z_i \in \mathcal{D}$ affect validation loss by measuring their gradient alignment with reducing validation loss on $\mathcal{V}_k$, which directly impacts task-specific evaluation scores. Denote a training data point as $\boldsymbol{z}$ and a validation data point as $\boldsymbol{z}'$ from validation set $\mathcal{V}_k$ for task T_k When training with SGD and batch size 1, using data point $\boldsymbol{z}$ at timestep t leads to a model update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \ell(\boldsymbol{z}; \boldsymbol{\theta}_t)$, where η_t is the learning rate. To reduce the computational cost, we use the first-order Taylor expansion to estimate the loss on a given validation data point $\boldsymbol{z}'$ at time step $t+1$ as:

$$\ell(\boldsymbol{z}'; \boldsymbol{\theta}_{t+1}) \approx \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t) + \langle \nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle.$$

The influence of a training data point $\boldsymbol{z}$ on a validation data point $\boldsymbol{z}'$ is:

$$\begin{aligned} \mathcal{I}_t(\boldsymbol{z} \to \boldsymbol{z}') &= \ell(\boldsymbol{z}'; \boldsymbol{\theta}_{t+1}) - \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t) \\ &\approx -\eta_t \langle \nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t), \nabla \ell(\boldsymbol{z}; \boldsymbol{\theta}_t) \rangle, \end{aligned}$$

which we refer to as an influence score. For our case of visual instruction tuning [23] which uses single-epoch training, the influence score is:

$$\mathcal{I}_{\text{single}}(\boldsymbol{z} \to \boldsymbol{z}') = \eta \langle \nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t), \nabla \ell(\boldsymbol{z}; \boldsymbol{\theta}_t) \rangle,$$

where η is the learning rate used during training. This can be generalized to multi-epoch training by accumulating the influence scores over all n epochs:

$$\mathcal{I}_{\text{total}}(\boldsymbol{z} \to \boldsymbol{z}') = \sum_{i=1}^{n} \eta_i \langle \nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_i), \nabla \ell(\boldsymbol{z}; \boldsymbol{\theta}_i) \rangle,$$

where η_i is the learning rate used during the ith epoch. The gradient-based selection approach selects training samples $\{\boldsymbol{z}\}$ that maximize the gradient inner product $\langle \nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_t), \nabla \ell(\boldsymbol{z}; \boldsymbol{\theta}_t) \rangle$[2] through a greedy, first-order approximation, which leads to larger reductions in validation loss for point $\boldsymbol{z}'$. While this approach omits second-order terms compared to influence functions [17], it provides an efficient approximation for ranking the impact of training samples [11, 42].

2.3. Specialist: Individual Task Influence Ranking

To rank the influence of training data for each target task, we compute the influence score of each training data point on a small validation set that represents the target task distribution. Following LESS [42], the process involves three steps: (1) training the model on 5% randomly selected data as a lightweight warm-up phase to initialize basic visual

[2] In practice, we use cosine similarity instead of direct inner products to avoid biasing selection toward shorter sequences, since gradient norms tend to be inversely correlated with sequence length as noted in [42].

instruction-following capabilities, (2) computing gradients for training and validation data and compressing the gradients via random projection, and (3) computing the influence score to quantify the impact of each training data on validation set.

Step 1: Warm-up Training. Following LESS [42], we first perform a brief warmup phase using Low-Rank Adaptation (LoRA) [12] on a small random subset $\mathcal{D}_{\text{warmup}} \subset \mathcal{D}$ (5% of the training data): $f_{\text{warmup}} = \text{LoRA}(f_{\text{base}}, \mathcal{D}_{\text{warmup}})$. This warmup phase allows the model to develop basic instruction-following capabilities.

Step 2: Gradient Computation. For each training data $\boldsymbol{z}_i \in \mathcal{D}$ and validation data $\boldsymbol{z}'_j \in \mathcal{D}^k_{\text{val}}$ from $\mathcal{T}_k$, we compute their gradients with respect to f_{warmup} parameters θ_w:

$$g_i = \nabla_{\theta_w} \mathcal{L}(f_{\text{warmup}}(\boldsymbol{z}_i), \boldsymbol{y}_i), \quad g'_j = \nabla_{\theta_w} \mathcal{L}(f_{\text{warmup}}(\boldsymbol{z}'_j), \boldsymbol{y}'_j)$$

where y_i and y'_j are the targets for z_i and z'_j, respectively. In order to reduce computational and storage overhead, we apply random projection to the gradient feature: $\tilde{g}_i = Rg_i$ and $\tilde{g}'_j = R'g'_j$, where $R, R' \in \mathbb{R}^{d' \times d}$ is a random projection matrix with $d' \ll d$ that preserves inner products with high probability [14]. We further normalize the projected gradient features, $\tilde{g}_i = \frac{\tilde{g}_i}{\|\tilde{g}_i\|_2}, \tilde{g}'_j = \frac{\tilde{g}'_j}{\|\tilde{g}'_j\|_2}$ to prevent bias from sequence length differences [42].

Step 3: Influence Matrix Computation. We compute the influence matrix $I \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}^k_{\text{val}}|}$ where each entry $I_{ij} = \langle \tilde{g}_i, \tilde{g}'_j \rangle$ is the influence of training data $\boldsymbol{z}_i$ on validation data $\boldsymbol{z}'_j$, and the average influence of training data $\boldsymbol{z}_i$ on the target task k is then computed as:

$$\bar{I}_k(\boldsymbol{z}_i) = \frac{1}{|\mathcal{D}^k_{\text{val}}|} \sum_{j=1}^{|\mathcal{D}^k_{\text{val}}|} I_{ij}. \tag{3}$$

This influence estimation process provides a task-specific ranking for the training set $\mathcal{D}$ with respect to task $\mathcal{T}_k$, where a higher influence score $\bar{I}_i$ suggests a higher influence for the target task. However, recall that our ultimate goal is to select a compact subset $\mathcal{S} \subset \mathcal{D}$ of size $M \ll N$ that maximizes $\sum_{k=1}^{K} \text{Rel}(f_{\mathcal{S}}, T_k)/K$ across all tasks. We address this disconnection between task-specific rankings and overall dataset optimization by proposing a voting-based generalist approach; this approach aggregates influence scores for each target task to identify the most broadly impactful training data.

2.4. Generalist: Cross-Task Influence Consensus

Our consensus-based voting strategy identifies training samples that consistently show a high influence score across various tasks. For each task k, we set a task-specific threshold τ_k at the $1-p$-th percentile of the influence score distribution within that task, where p represents the proportion of top samples to consider ($p = 0.2$ in our main experiments). A sample $\boldsymbol{z}_i$ is considered important for task k if its influence score $I_k(\boldsymbol{z}_i) \geq \tau_k$. Formally, we define the voting-based influence score for each sample z_i on task k as:

$$I_{\text{vote}}(\boldsymbol{z}_i) = \sum_{k=1}^{K} \mathbb{1}[I_k(\boldsymbol{z}_i) \geq \tau_k], \tag{4}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function, and $I_k(\boldsymbol{z}_i)$ represents the influence score of sample $\boldsymbol{z}_i$ for task k. The final score for each sample is the total number of votes it receives across all tasks, ranging from 0 (indicating no importance for any task) to K (indicating consistent importance for all tasks). Influence Consensus identifies a unique subset of samples that are consistently influential across multiple tasks, rather than those that might be highly influential in only a few tasks. After computing the aggregated influence scores, we rank the training samples and select the top $p\%$ as final subset for model training.

While directly combining validation data from all target tasks and computing a single influence matrix (Direct-merge Selection, §3.5) might seem intuitive, we adopt this consensus-based approach for three reasons. First, specialist selection provides a practical performance upper bound, as the performance achieved by selecting the top $p\%$ of data specifically for a single task represents the theoretical maximum achievable with that data budget, serving as a critical baseline for evaluating generalist selection trade-offs (§3.5). Second, this approach is scalable and naturally supports continual learning scenarios, as influence matrices for new tasks or benchmarks can be computed independently without recomputing training gradients, thereby reducing computational costs. Third, computing influence matrices independently for different tasks give us better understanding of cross-task influence dynamics. Our analysis (§3.3) reveals that the influence ranking of training data vary significantly across different tasks, with highly influential samples for one task often showing limited impact on others. This further provides insights to design more effective consensus mechanisms for generalist selection while maintaining computational tractability.

3. Experiments

In this section, we first discuss our experiment setup and evaluation benchmarks (§3.1). Then we compare it with the latest state-of-the-art methods (§3.2). Next, we provide detailed analysis of our method's behavior and effectiveness across different dimensions (§3.3). We further evaluate the transferability of our selection approach across both unseen tasks and model architectures (§3.4). Finally, we conduct ablation studies to understand the contribution of our approach (§3.5).

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-W Bench	Rel. (%)
Full	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	100
Random	75.7	58.9	44.3	68.5	55.3	84.7	1483.0	62.2	54.8	65.0	95.8
CLIP-Score [34]	73.4	51.4	43.0	65.0	54.7	85.3	1331.6	55.2	52.0	66.2	91.2
EL2N [32]	76.2	58.7	43.7	65.5	53.0	84.3	1439.5	53.2	47.4	64.9	92.0
Perplexity [27]	75.8	57.0	47.8	65.1	52.8	82.6	1341.4	52.0	45.8	68.3	91.6
SemDeDup [1]	74.2	54.5	46.9	65.8	55.5	84.7	1376.9	52.2	48.5	**70.0**	92.6
D2-Pruning [26]	73.0	58.4	41.9	69.3	51.8	85.7	1391.2	**65.7**	**57.6**	63.9	94.8
Self-Sup [37]	74.9	59.5	46.0	67.8	49.3	83.5	1335.9	61.4	53.8	63.3	93.4
Self-Filter [5]	73.7	58.3	**53.2**	61.4	52.9	83.8	1306.2	48.8	45.3	64.9	90.9
COINCIDE [18]	**76.5**	59.8	46.8	69.2	**55.6**	86.1	**1495.6**	63.1	54.5	67.3	97.4
ICONS (ours)	76.3	**60.7**	50.1	**70.8**	**55.6**	**87.5**	1485.7	63.1	55.8	66.1	**98.6**

Table 1. **Baseline Comparisons.** Performance comparison of different data selection approaches when trained on 20% of the LLaVA-665K dataset. Models are evaluated across diverse multimodal benchmarks, with tasks ranging from visual question answering (VQAv2, GQA) to multilingual understanding (MMBench). The best and second best results for each benchmark are shown in **bold** and underlined, respectively. Our method ICONS achieves the highest overall relative performance (98.6%), consistently outperforming existing approaches including COINCIDE (97.4%) and D2-Pruning (94.8%), while methods like EL2N, Perplexity, and CLIP-Score show limited effectiveness with relative performance around 91-92%.

3.1. Evaluation Test-Bed

Dataset & Model. We test ICONS on visual instruction tuning dataset LLaVA-665K [22]. We use `LLaVA-v1.5-7b-lora` model [22] with a default size of 7B parameters unless otherwise specified. In all experiments, we train the models for one epoch following the official finetuning hyperparameters, which significantly reduces computational overhead while maintaining selection quality.

Target Tasks. We evaluate the effectiveness of ICONS with a wide range of multimodal benchmarks (Tab. 2) that test different capabilities of vision-language models. The benchmarks include: 1) Multiple-choice understanding: MMBench [47] and MME [7], which assess the model's ability to select correct answers from given options; 2) Visual question answering: VQAv2 [9], GQA [13], and VizWiz [10], which test basic visual reasoning; 3) Text understanding in images: TextVQA [36], which evaluates OCR capabilities; 4) Scientific reasoning: ScienceQA [25], which tests knowledge-grounded question answering; 5) Open-ended generation: LLaVA-W Bench [23], which evaluate free-form response generation; 6) Factual consistency: POPE [21], which measures hallucination.

	MME	POPE	SQA-I	MMBench en	MMBench cn
$\lvert\mathcal{D}_{\text{val}}\rvert$	986	500	424	1,164	1,164
$\lvert\mathcal{D}_{\text{test}}\rvert$	2,374	8,910	4,241	1,784	1,784
Task Type	Y/N	Y/N	MCQ	MCQ	MCQ
	VQAv2	**GQA**	**VizWiz**	**TextVQA**	**LLaVA-W**
$\lvert\mathcal{D}_{\text{val}}\rvert$	1,000	398	8,000	84	84
$\lvert\mathcal{D}_{\text{test}}\rvert$	36,807	12,578	8,000	5,000	84
Task Type	VQA	VQA	VQA	VQA	VQA

Table 2. **Statistics of Target Tasks.** Our target tasks include diverse benchmarks and answer formats, covering different vision-language capabilities. Task types include Multiple-Choice Questions (**MCQ**), Visual Question Answering (**VQA**), and Yes/No Questions (**Y/N**).

3.2. Comparisons with Baselines

We compare our ICONS against several baseline approaches, including random selection. We evaluate CLIP-Score [34] for measuring image-text alignment, EL2N [32] based on embedding L2 norms, and Perplexity [27] using language model scores. We also compare against SemDeDup [1] for semantic deduplication and D2-Pruning [26] for distribution-aware pruning. Additional baselines include Self-Sup [37] leveraging self-supervised signals, while Self-Filter [5] and COINCIDE [18] are the approaches specifically designed for vision-language data selection. All methods are evaluated using a 20% sampling ratio of the original dataset, with results presented in Tab. 1. We reference the baseline results from COINCIDE [18].

We compare our ICONS against existing data selection approaches in Tab. 1. Our method achieves the best overall performance with 98.6% Rel., outperforming all baselines with LLaVA-ICONS-133K, 20% of the training data. Remarkably, we achieve comparable or better performance than training with the entire LLaVA-665K on several tasks: SQA-I (70.8 vs. 68.4), MME (1485.7 vs. 1476.9), VizWiz (50.1 vs. 47.8) and POPE (87.5 vs. 86.4). Among the baselines, COINCIDE achieves strong performance (97.4% Rel.), though it falls short of ICONS

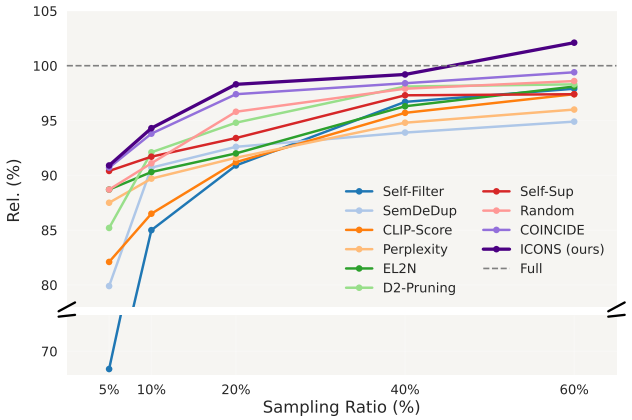


Figure 3. **Performance Comparison at Different Selection Ratios.** Here we show the average relative performance comparisons of various data selection methods at different selection ratios. Our ICONS consistently outperforms all baseline approaches across different selection ratios, achieving 99.2% relative performance with 40% of the data, and remarkably exceeding 102% at 60% selection ratio. This shows that our selected subset is not only more efficient but can be even more effective than the full dataset for model training.

Task	Full	Specialist	Generalist	Delta (%)
VQAv2 [9]	79.1	77.1	76.3	1.04
GQA [13]	63.0	61.1	60.7	0.65
VizWiz [10]	47.8	53.1	50.1	5.65
SQA-I [25]	68.4	69.8	70.8	-1.43
TextVQA [36]	58.2	55.7	55.6	0.18
POPE [21]	86.4	86.6	87.5	-1.04
MME [7]	1476.9	1506.1	1485.7	1.35
MMBench (en) [47]	66.1	66.0	63.1	4.34
MMBench (cn) [47]	58.9	56.4	55.8	1.06
LLaVA-W Bench [23]	67.9	67.1	66.1	1.49

Table 3. **Single-task Selection (Specialist) vs. Consensus-aware Multi-task Selection (Generalist).** The single-task data selection approach selects 20% of data per task, while our consensus-aware multi-task data selection approach selects a total of 20% data across all tasks. The Delta (%) column represents the relative drop in performance from specialist to generalist.

on key tasks. Other methods like EL2N, Perplexity, SemDeDup and CLIP-Score are less effective, with relative scores around 91-92%, showing limitations in preserving performance.

3.3. Analysis

From Specialist to Generalist. In order to understand the effectiveness of the intermediate task-specific influence matrices we obtained from the specialist stage, we further conduct experiments on selecting 20% of data for individual tasks. As shown in Tab. 3, using 20% task-specific data, it achieves comparable or even superior performance

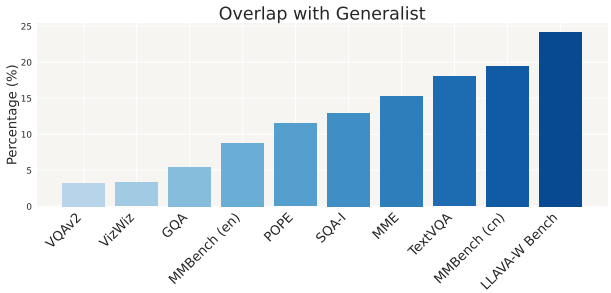


Figure 4. **Data Overlap between Specialists and Generalist Selection.** It shows varying degrees of overlap, ranging from 3.27% (VQAv2 [9]) to 24.21% (LLAVA-W Bench [23]).

to full data training in several cases (e.g., MME from 1476.9 to 1506.1 and SQA-I from 68.4 to 69.8). Then, through our influence consensus mechanism in the generalist stage, we identify consistently influential cross-tasks samples and select a 20% subset of the total data. This generalist approach maintains remarkably strong performance, with an average drop of 1.33% across most tasks compared to specialist baselines.

As visualized in Fig. 4, the overlap between specialist and generalist selections varies significantly across tasks, ranging from minimal overlap in simpler tasks like VQAv2 [9] (3.27%) and VizWiz [10] (3.28%) to substantial agreement in more complex tasks like LLAVA-W Bench [23] (24.21%). Notably, some tasks even show performance improvements under the generalist selection, such as SQA-I [25] (improving by 1.43%) and POPE [21] (improving by 1.04%). This empirically validates our consensus-based approach: by first understanding task-specific influence patterns and then building consensus across tasks, we can effectively identify a compact, universal training set that maintains strong performance across diverse vision-language tasks while significantly reducing the required training data.

Different Selection Ratio. We evaluate our ICONS against baseline methods across different selection ratios, ranging from 5% to 60% of the training data. As shown in Fig. 3, our results reveal several key patterns: First, ICONS shows particularly strong performance in the low-selection regime (5-20%), where identifying the most influential samples is crucial. Another notable observation is that as the selection ratio increases, the performance gap between different methods gradually narrows. This convergence pattern is expected, as larger sample sizes naturally capture more of the dataset's diversity and information. Despite this convergence trend, ICONS consistently maintains its performance edge across all selection ratios, and remarkably reaching above 100% at 60% selection ratio. One hypothesis is that our method can also effectively filter out potentially harmful or noisy training samples that might negatively impact

model training, thereby surpassing the full training performance.

Divergent Cross-Task Influence Patterns. As shown in Fig. 5a, the pairwise overlap heatmap reveals significant variation in training data influence rankings across different tasks. High overlap, such as between VQAv2 and VizWiz (49.0%) or POPE and GQA (60.2%), suggests that certain samples are beneficial across similar tasks. However, low overlap, like the 3.3% between MMBench (en) and GQA, highlights that highly influential samples for one task may have limited impact on others. Even closely related tasks, such as MMBench in different languages (English and Chinese), share 67.4% of influential samples. These findings empirically demonstrate significant overlap in influential samples across tasks, motivating the use of influence consensus for effective multi-task data selection.

3.4. Transferability

We evaluate the transferability of our data selection approach across two dimensions: unseen-task generalization and cross-architecture-scale generalization.

Unseen-task Generalization. Here we measure the performance on entirely unseen benchmarks that were not considered during our data selection process. As shown in Tab. 4, we test on diverse evaluation sets including MMVet [45], NaturalBench [19], AI2D [16], ChartQA [28], DocVQA [29], InfoVQA [30], RealWorldQA [41] and CMMMU [46]. LLaVA-ICONS-133K shows strong transferability, maintaining 95.5-113.9% (Rel.) compared to full dataset training. Importantly, LLaVA-ICONS-133K significantly outperforms random selection across all benchmarks - for example, improving performance on ChartQA (17.1 vs. 15.1), MMVet (29.7 vs. 27.6), and NaturalBench (12.8 vs. 11.1). This consistent improvement over random selection, combined with the strong relative performance to full training. This suggests that our selection approach successfully captures fundamental visual-language understanding capabilities that transfer well across different task formats and domains, and consistently outperforms random selection by a significant margin. Notably, we observe improvements on InfoVQA (103.8%), NaturalBench (105.5%), RealWorldQA (104.4%), and CMMMU (113.9%), suggesting that, in some cases, training on LLaVA-ICONS-133K may even outperform training on the full dataset, despite these tasks not being included in the selection process.

Cross Architecture Scale Generalization. We further conduct experiments on cross architecture scale generalization to evaluate the transferability of our selected data across different model scales. While our subset was initially selected using `LLaVA-1.5-7B` as the base model, we investigate whether these same examples remain effective for training larger architectures like `LLaVA-1.5-13B`. This tests whether our selection criteria identify universally valuable training examples rather than model-specific patterns. Our results in Tab. 5 show cross architecture scale generalization, with the 7B-selected achieving 98.1% relative performance compared to full training, suggesting that our selected subset captures fundamental visual-language understanding capabilities that scale effectively across model sizes.

3.5. Ablation Studies

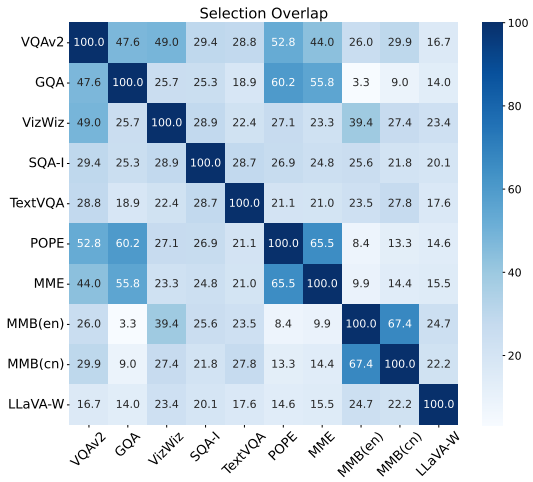
Different Aggregation Approaches. To understand how different methods of combining task-specific influence scores affect the performance, we compare our proposed majority voting approach (**Vote**) with four aggregation approaches: (1) **Mean**, which averages scores across tasks, (2) **Max**, which selects the highest score per sample, (3) **Rank**, which averages the relative ranking of samples within each task, (4) **Norm**, which normalize the scores before averaging:

$$I_{\text{mean}}(\boldsymbol{z}_i) = \frac{1}{K}\sum_{k=1}^{K} I_k(\boldsymbol{z}_i) \qquad I_{\text{norm}}(\boldsymbol{z}_i) = \frac{1}{K}\sum_{k=1}^{K} \frac{I_k(\boldsymbol{z}_i) - \mu_k}{\sigma_k}$$

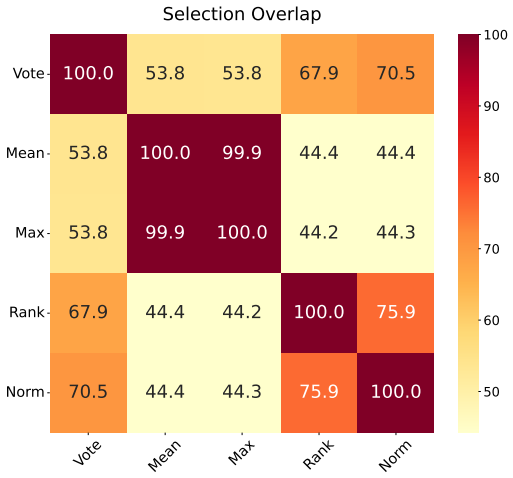
$$I_{\text{max}}(\boldsymbol{z}_i) = \max_{k=1,\ldots,K} I_k(\boldsymbol{z}_i) \qquad I_{\text{rank}}(\boldsymbol{z}_i) = \frac{1}{K}\sum_{k=1}^{K} \text{rank}(I_k(\boldsymbol{z}_i))$$

Results in Tab. 6 show that our Vote consistently outperforms alternatives across most tasks, achieving the highest overall relative performance (**Rel.**) of 98.6%. While methods like Rank perform well for MME (1490.0 vs. 1485.7 with Norm), they underperform overall. To better understand the relationships between these aggregation approaches, we visualize their pairwise selection overlap in Fig. 5b. Specifically, the **Mean** and **Max** approaches show an extremely high overlap of 99.9%, suggesting they tend to select very similar influential samples. In contrast, our **Vote**-based approach exhibits a more moderate overlap with other methods, ranging from 53.8% to 70.5%, potentially those that are influential across a broader variety of tasks rather than overfitting to specific tasks.

Consensus-aware Selection vs. Direct-merge Selection. To validate our choice of a consensus-aware approach (first computing task-specific influence matrices and then aggregating results) versus a direct-merge approach (directly computing influence using a combined validation set), we compare the performance of both approaches. In the direct-merge approach, we combine validation sets from all tasks into a single pool, denoted as $\{\boldsymbol{z}'\}_{\text{merged}}$, and compute influence scores directly against this combined set: $\mathcal{I}_{\text{direct}}(\boldsymbol{z} \rightarrow \{\boldsymbol{z}'\}_{\text{merged}})$. As described in Tab. 7, our consensus-aware approach consistently outperforms direct-merge selection across all benchmarks, with the largest gains observed on complex tasks like MME [7].



(a) **Pairwise Overlap Heatmap between Specialist Benchmarks.** The values are the percentage of selected overlapping samples across different benchmarks, with high overlap seen between POPE and GQA (60.2%) and between VQAv2 and VizWiz (49.0%). Lower overlap, such as 3.3% between MMBench(en) and GQA, shows more distinct selected sample sets.



(b) **Pairwise Overlap Heatmap between Different Aggregation Approaches.** The values are the percentage of overlapping selected samples between each pair of aggregation approaches. Our Vote-based influence consensus approach shows moderate overlap (53.8%-70.5%) with other methods, while Mean and Max selections are highly correlated (99.9%).

Figure 5. **Pairwise Overlap Heatmaps between Different Aggregation Approaches and Benchmarks.** (a) compares the overlap between different specialist benchmarks, while (b) shows overlap between various aggregation methods.

	AI2D	ChartQA	DocVQA	InfoVQA	MMVet	Naturalbench	RealworldQA	CMMMU	Rel. (%)
Full	55.4	17.5	28.9	26.5	31.1	12.4	52.4	22.1	100.0
Random	50.2	15.1	25.2	24.3	27.6	11.1	49.8	21.9	91.6
LLaVA-ICONS-133K	53.9	17.1	27.9	27.5	29.7	12.8	55.0	25.2	98.7
Per-task Rel. (%)	97.3	97.7	96.5	103.8	95.5	103.2	104.4	114.0	-

Table 4. **Unseen-Task Generalization.** Performance comparison on unseen benchmarks when trained on the selected subset. The relative performance ratios show that our selection approach successfully captures fundamental visual-language understanding capabilities that transfer well across different task formats and domains. Notably, we observe improvements on InfoVQA [30] (103.8%), MMVet [45] (95.5%), NaturalBench [19] (103.2%), RealWorldQA [41] (104.4%), and CMMMU [46] (114.0%), suggesting that in some cases, training on our LLaVA-ICONS-133K may even outperform using the full dataset.

	VQAv2	GQA	Vizwiz	SQA-I	TextVQA	POPE	MME	MMBench	MMBench(cn)	LLAVA-W	Rel. (%)
Full	80.0	63.3	58.9	71.2	60.2	86.7	1541.7	68.5	61.5	69.5	100.0
Random	77.3	60.7	57.6	69.1	56.8	82.9	1517.2	63.2	56.3	67.5	95.7
7B-selected	78.8	60.4	57.4	70.4	58.3	84.3	1527.5	64.9	59.7	68.2	97.3
13B-selected	78.9	61.2	57.5	71.3	58.4	85.9	1535.2	66.1	59.8	68.8	98.1

Table 5. **Cross Architecture Scale Generalization.** Our LLaVA-ICONS-133K selected via `LLaVA-v1.5-7b-lora` (7B-selected) shows strong cross-architecture transferability, achieving competitive performance when trained with data selected via `LLaVA-v1.5-13b-lora` (13B-selected). Specifically, while the 7B-selected achieves 97.3% relative performance, the 13B-selected model reaches 98.1%, showing that our selected subset's effectiveness scales well to larger architectures. Both significantly outperform random selection (95.7%), with the 13B-selected option showing particular strength in complex reasoning tasks like MMBench and POPE. This consistency across model scales highlights that our data selection approach captures fundamental visual-language understanding patterns that generalize well across different model architecture scales.

A potential reason could be the direct-merge approach does not account for the different vision-language task difficulties or validation set sizes across the target tasks. By pooling all validation examples, it effectively treats each validation data from all tasks equally, which can lead to the influence scores being dominated by tasks with more samples or with higher influence scores on average. The task-aware voting-based influence consensus mechanism of our direct-merge approach are more robust to these issues. Additionally, our approach is more scalable, as it avoids com-

Task	Mean	Max	Rank	Norm	Vote (ours)
VQAv2 [9]	75.7	75.2	74.9	75.1	**76.3**
GQA [13]	59.6	59.8	58.6	60.1	**60.7**
VizWiz [10]	47.9	48.1	40.5	46.4	**50.1**
SQA-I [25]	65.5	66.2	69.8	69.8	**70.8**
TextVQA [36]	55.5	55.5	55.2	54.5	**55.6**
POPE [21]	86.0	85.5	85.6	85.6	**87.5**
MME [7]	1422.1	1470.7	**1490.0**	1482.6	1485.7
MMBench (en) [47]	59.0	58.3	59.0	58.9	**63.1**
MMBench (cn) [47]	51.0	51.8	50.8	52.5	**55.8**
LLaVA-W Bench [23]	66.2	66.2	**66.4**	66.3	66.1
Rel.(%)	96.4	96.1	95.9	96.8	**98.6**

Table 6. **Comparison of Different Aggregation Approaches.** Performance comparison of different methods for combining task-specific influence scores, using 20% of training data. All scores are relative to full dataset training.

Task	Full	Direct merge	Consensus aware (ours)	Delta (%)
VQAv2 [9]	79.1	76.1	76.3	0.26
GQA [13]	63.0	59.4	60.7	2.19
VizWiz [10]	47.8	46.1	50.1	8.67
SQA-I [25]	68.4	68.7	70.8	3.06
TextVQA [36]	58.2	54.1	55.6	2.77
POPE [21]	86.4	85.1	87.5	2.82
MME [7]	1476.9	1419.2	1485.7	4.69
MMBench (en) [47]	66.1	61.9	63.1	1.94
MMBench (cn) [47]	58.9	50.3	55.8	10.94
LLaVA-W Bench [23]	67.9	65.2	66.1	1.38
Rel. (%)	100	94.7	98.6	-

Table 7. **Comparison of Consensus-aware vs. Direct-merge Selection.** Performance comparison across different benchmarks using 20% of training data. Direct-merge computes influence directly on combined validation sets, while consensus-aware refers to our specialist-to-generalist approach. The Delta (%) column represents the relative improvement from Direct-merge to Consensus-aware. Consensus-aware outperforms Direct-merge across all benchmarks, while Direct-merge notably underperforms even the random baseline.

puting a single, large influence matrix across validation data from all target tasks. Adding new tasks only requires computing influence scores for those specific tasks and updating the voting mechanism, while direct-merge selection would need to recompute the entire large influence matrix which makes it less practical.

4. Related Work

[Olga: I think we need prior work (including corset selection and such, active learning, ...) from the pre-visual instruction tuning days. There is lots of literature on data selection, whereas this makes is seem like it's a novel problem.][Xindi: fix2]

Visual Instruction Tuning. Multimodal large language models (MLLMs), e.g., Flamingo [2], LLaVA [23], BLIP2 [20], and Cambrian [39], enhance the capabilities of large language models (LLMs) on various multimodal tasks. A key component in advancing MLLMs is visual instruction tuning [23], a training process that enables these models to interpret and follow instructions within a vision-language context, transforming them into versatile multimodal assistants. This tuning process not only improves the models' instruction-following abilities but also aligns their outputs more closely with user expectations, thus enhancing their utility in practical applications [23].

Data Selection. Data selection methods can be categorized based on the types of information they utilize for selection. Representation-based approaches [1, 18] leverage neural embeddings to capture data representations. Loss trajectory-based methods [31] prioritize data points that contribute most significantly to reducing generalization error over training. Gradient-based techniques [6, 32, 42] select data based on gradient information. Recent work has explored various approaches to select optimal visual instruction tuning datasets. Concurrent work TIVE [24] employs gradient-based selection to identify representative instances. TIVE assumes that the number of specialist data should be proportional to task difficulty and thus samples specialist data based on an estimation of task difficulty. Our method does not rely on this assumption – we directly select samples that benefit the greatest number of tasks. COINCIDE [18] clusters data based on representations associated with concept-skill compositions.

Our work follows targeted instruction tuning selection approach LESS [42] to utilize gradient information to calculate the specialist influence (i.e., the influence on a specific task) and extends it to general scenarios by aggregating information from various tasks and selecting data for multiple downstream tasks via majority voting.

5. Conclusion

In this work, we introduce ICONS, a simple yet effective influence consensus-based approach for visual instruction tuning data selection. Our two-stage specialist-to-generalist method, first computes task-specific influence matrices and then select a subset by prioritizing universally influential samples through a cross-task voting mechanism. Through extensive experiments, we show that ICONS achieving 98.6% of full dataset performance using only 20% of the data and it consistently outperform baseline methods across different selection ratios. Additionally, we release LLaVA-ICONS-133K, a 20% subset of LLaVA-665K dataset, that not only maintains strong performance across diverse vision-language tasks but also shows transferability to unseen tasks. We hope our work inspires further exploration into data-efficient methods for vision-language models across diverse applications.

Limitations. Our approach primarily faces one practical limitation: computing gradients for large training datasets is computationally expensive. This computational overhead could potentially constrain the method's applicability when working with extremely large-scale datasets. To support broader research community, we release our LLaVA-ICONS-133K dataset to help research iteration and model development under resource-constrained settings.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 2107048 and No.2112562. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work is also supported by the Singapore National Research Foundation and the National AI Group in the Singapore Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-001). We thank many people for their helpful discussion and feedback, listed in alphabetical order by last name: Allison Chen, Hamish Ivison, Carlos E. Jimenez, Polina Kirichenko, Jaewoo Lee, Tiffany Ling, Zhiqiu Lin, Ethan Tseng, Shengbang Tong, Justin Wang, Zirui Wang.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 5, 9

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 9

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[4] Alexander Bukharin and Tuo Zhao. Data diversity matters for robust instruction tuning. *arXiv preprint arXiv:2311.14736*, 2023. 1

[5] Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*, 2024. 5

[6] Zhiwei Deng, Tao Li, and Yang Li. Influential language data selection via gradient trajectory pursuit. 2024. 1, 9

[7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. corr abs/2306.13394 (2023), 2023. 5, 6, 7, 9, 2

[8] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5, 6, 9, 2

[10] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5, 6, 9, 2

[11] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113 (5):2351–2403, 2024. 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5, 6, 9, 2

[14] William B Johnson. Extensions of lipshitz mapping into hilbert space. In *Conference modern analysis and probability, 1984*, pages 189–206, 1984. 4

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 7

[17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. 1, 3

[18] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*, 2024. 2, 5, 9

[19] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 7, 8

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 9

[21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5, 6, 9, 2

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 5

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3, 5, 6, 9

[24] Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024. 9

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5, 6, 9, 2

[26] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*, 2023. 5

[27] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023. 5

[28] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 7

[29] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 7

[30] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 7, 8

[31] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR, 2022. 9

[32] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021. 5, 9

[33] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[35] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 1

[36] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5, 6, 9, 2

[37] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022. 5

[38] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13693–13696, 2020. 1

[39] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 9, 2

[40] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022. 1

[41] x.ai. Introducing Grok 1.5v: The Latest Advancement in AI, 2024. [Online; accessed 14-November-2024]. 7, 8

[42] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024. 1, 2, 3, 4, 9

[43] Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*, 2022. 1

[44] Simon Yu, Liangyu Chen, Sara Ahmadian, and Marzieh Fadaee. Diversify and conquer: Diversity-centric data selection with iterative refinement. *arXiv preprint arXiv:2409.11378*, 2024. 1

[45] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 7, 8

[46] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu,

Shuyue Guo, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024. 7, 8

[47] Yuanhan Zhang Bo Li-Songyang Zhang, Wangbo Zhao Yike Yuan Jiaqi Wang, Conghui He Ziwei Liu Kai Chen, Dahua Lin Yuan Liu, and Haodong Duan. Mmbench: Is your multi-modal model an all-around player. *arXiv preprint arXiv:2307.06281*, 2, 2023. 5, 6, 9, 2

Appendix for ICONS: Influence Consensus for Vision-Language Data Selection

In this supplement, we provide additional experiment details and ablation studies (§A), additional analysis (§B). We detail our algorithmic framework (§C) and discuss future research directions (§D). Finally, we include visualization of the selected data (§E) to better understand data samples with high influence.

A. Additional Experiment Details & Ablations

Projection Dimension. We primarily set the projection dimension to 5120, reducing features from 338.7M to 5120 dimensions. The choice of 5120 was empirically validated for its trade-off between effective capturing gradient representation and maintaining a manageable parameter space. Our `LLaVA-v1.5-7b-lora` architecture includes a total of 7.4B parameters, with 338.7M parameters being trainable after LoRA adaptation, accounting for approximately 4.58% of the total parameter count. We further ablate different projection dimensions (1024, 2560, 5120, and 10240), with results provided in Fig. 6.

Warm-up Ratio. To initiate training, we use a warm-up set including 5% of the total training data. We conducted ablation studies to evaluate the impact of varying warm-up ratios (5%, 10%, 20%, and 100%) on selection performance, as shown in Fig. 7. Our experiments reveal that increasing the warm-up data size does not lead to performance improvements. Surprisingly, models trained with smaller warm-up ratios (5-20%) consistently outperform those trained with the full dataset (100%). Specifically, the 5% warm-up ratio achieves the best performance at 98.6%, while using the complete dataset results in a performance drop to 97.8%. This finding suggests that a small subset of training data is sufficient and even beneficial for model initialization, and potentially gives better signals in the early training stages.

B. Additional Analysis

Consistency Analysis. To evaluate the consistency of our approach, we conduct three independent runs of the experiment. As shown in Fig. 8, our method demonstrates high consistency across different runs, achieving 98.6±1.2% Rel., which shows a notable improvement over the random baseline, which achieves 95.8±2.7%. The lower standard deviation in our results (1.2% vs 2.7%) further indicates that our approach produces more stable and reliable outcomes compared to the random baseline.

B.1. Computational Complexity

Computing gradient-based influence requires significant computational resources. In the specialist stage, the complexity scales with both the dataset size $|\mathcal{D}|$ and the gradient dimension d. This stage consists of three steps. First, the warm-up training has a complexity of $\mathcal{O}(|\mathcal{D}_{\text{warmup}}|)$. Second, the gradient computation stage has a computational complexity of $\mathcal{O}(|\mathcal{D}| + |\mathcal{D}_{\text{val}}|)$ for forward and backward

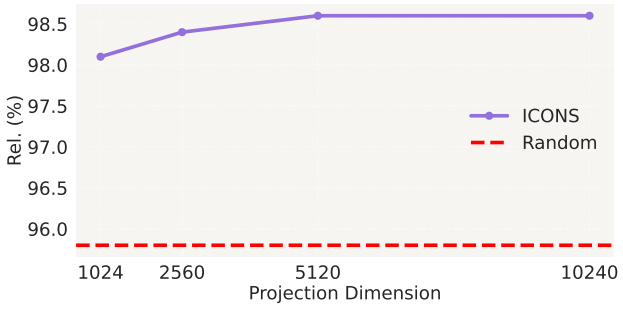


Figure 6. **Projection Dimension Ablation.** We show the performance of ICONS at different projected dimensions (1024, 2560, 5120, 10240), compared to the random baseline. The performance increases with the projected dimension and reaches a plateau around dimension 5120.

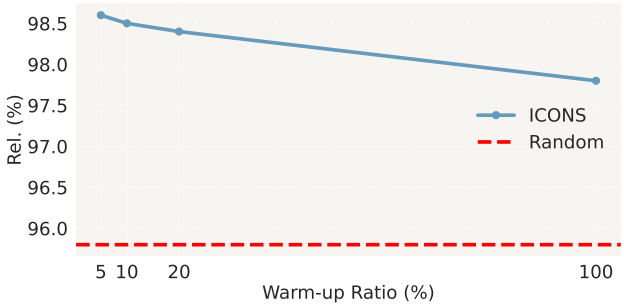


Figure 7. **Warm-up Ratio Ablation.** The blue line represents ICONS performance across different warm-up ratios (5%, 10%, 20%, and 100%), while the red dashed line shows the random baseline performance. Results show that smaller warm-up ratios (5-20%) achieve better performance compared to using the full dataset (100%).

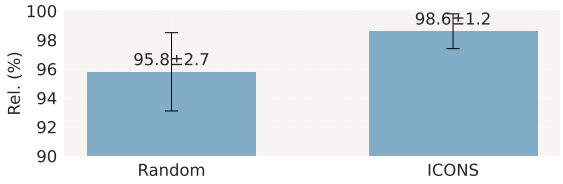


Figure 8. **Rel. (%) Across Runs.** We show the performance across three different runs for random selection and our ICONS. Our ICONS achieves 98.6±1.2% while the random baseline reaches 95.8±2.7%.

passes, with storage requirements of $\mathcal{O}(|\mathcal{D}| \cdot d + |\mathcal{D}_{\text{val}}| \cdot d)$ for the gradients. Third (and finally), the influence matrix computation requires $\mathcal{O}(|\mathcal{D}| \cdot |\mathcal{D}_{\text{val}}| \cdot d')$ compute cost, where d' is the reduced dimension after projection. The generalist stage, focusing on influence consensus across tasks, has lower computational requirements. It begins with threshold computation, requiring $\mathcal{O}(K \cdot |\mathcal{D}| \log |\mathcal{D}|)$ operations for sorting across K tasks. The voting process then takes $\mathcal{O}(K \cdot |\mathcal{D}|)$ compute, followed by a final selection step with complexity $\mathcal{O}(|\mathcal{D}| \log |\mathcal{D}|)$ for sorting the aggregated votes. Storage requirements for this stage are minimal, primarily needing space for the final selected subset.

In practice, for LLaVA-665K training data, the warmup training phase requires 0.75 hours using eight L40 GPUs. We parallelize the gradient computation across 100 A6000 GPUs, taking approximately one hour and requiring 103GB of total storage for the gradients. The influence consensus stage is notably efficient, completing in less than a minute on a single L40 GPU. While these computational demands are substantial, they represent front-loaded, one-time costs that can be used across multiple target tasks and model iterations. This makes our method extendable for new tasks, as the expensive training data gradient computations only need to be performed once.

B.2. Visual Dependency Influence Ranking

Recent work [39] has shown that vision-language tasks vary in their reliance on visual information: task like MMBench [47] depends heavily on visual grounding, while others like SQA-I [25] can be handled primarily through language, showing only a 5% drop in performance when visual input is removed [39]. To take visual dependency of training data into consideration, we further explored gradient-based Visual Dependency Score (VDS). For each data point, we calculate the gradient of the model's autoregressive cross-entropy loss with both the original image and a Gaussian noise image $I_{\text{noise}} \sim \mathcal{N}(0, 1)$, keeping the text input constant. This quantifies how much the visual component contributes to the model performance. We construct an adapted influence matrix: visual influence matrix $\mathcal{I}_{\text{VDS}} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}_{\text{val}}|}$, which quantifies the visual influence of each training sample $\boldsymbol{z}_i$ on each validation sample $\boldsymbol{z}'_j$ with respect to the model's gradient alignment and visual dependency. $\mathcal{I}_{\text{VDS}}$ is computed as:

$$\mathcal{I}_{\text{VDS},ij} = \langle \nabla_\theta \mathcal{L}(\boldsymbol{z}'_i), \nabla_\theta \mathcal{L}(\boldsymbol{x}_j, I_j) - \nabla_\theta \mathcal{L}(\boldsymbol{x}_j, I_{\text{noise}}) \rangle, \quad (5)$$

where $\nabla_\theta \mathcal{L}(\boldsymbol{x}_j, I_j)$ and $\nabla_\theta \mathcal{L}(\boldsymbol{x}_j, I_{\text{noise}})$ are the gradients computed with the original and Gaussian noise images, respectively. The visual influence matrix $\mathcal{I}_{\text{VDS}}$ provides insights into which training samples have the most influence on the validation samples from a visual perspective. This matrix can be used to further rank and select training data that are most impactful for tasks requiring strong visual grounding, ensuring that the selected subset effectively supports vision-dependent performance.

Task	w/o VDS	w/ VDS	Δ (%)
VQAv2 [9]	76.3	75.8	-0.66
GQA [13]	60.7	60.9	+0.33
VizWiz [10]	50.1	50.3	+0.40
SQA-I [25]	70.8	69.5	-1.84
TextVQA [36]	55.6	54.8	-1.44
POPE [21]	87.5	86.8	-0.80
MME [7]	1485.7	1489.3	+0.24
MMBench (en) [47]	63.1	64.3	+1.90
MMBench (cn) [47]	55.8	56.3	+0.90
LLaVA-W Bench [23]	66.1	67.9	+2.72

Table 8. **Impact of Visual Dependency Score (VDS) on Selection Performance.** Performance comparison between influence calculation with and without VDS. VDS shows strongest improvements on LLaVA-W Bench (+2.72%), MMBench (en) (+1.90%), and MMBench (cn) (+0.90%), while tasks like SQA-I (-1.84%), TextVQA (-1.44%), and POPE (-0.80%) show significant performance drops.

Our empirical results demonstrate that VDS-based data selection has varying effectiveness across different vision-language tasks (Tab. 8). The approach shows substantial improvements on tasks requiring strong visual understanding, such as open-ended generation (LLaVA-W Bench: +2.72%) and multiple-choice understanding (MMBench-EN: +1.90%, MMBench-CN: +0.90%). However, tasks that primarily rely on textual reasoning show decreased performance, including SQA-I (-1.84%) and TextVQA (-1.44%). These results align with and extend the findings in Cambrian [39], demonstrating that the effectiveness of VDS corresponds to a task's visual dependency - tasks that maintain performance without visual inputs show limited or negative impact from VDS-based selection, while visually-dependent tasks benefit significantly. This pattern suggests that VDS effectively identifies training samples where visual information plays an important role in model learning.

C. Algorithm Details

We provide detailed pseudocode for our two-stage ICONS framework. Stage 1 (specialist) computes task-specific influence scores through gradient-based analysis with efficient random projections. Stage 2 (generalist) implements our voting-based consensus mechanism to select samples that are influential across multiple tasks.

Algorithm 1 ICONS Stage 1: Specialist (Task-specific Influence Computation)

Require: Training dataset $D = \{(x_i, I_i, y_i)\}_{i=1}^{N}$, target tasks $T = \{T_1, ..., T_K\}$
Require: Warm-up ratio r (default 5%)
Ensure: Task-specific influence scores $\{\bar{I}_k\}_{k=1}^{K}$

for each task $T_k \in T$ **do**
  // Step 1: Warm-up Training
  Sample warm-up set $D_{\text{warmup}} \subset D$ of size $r|D|$
  $f_{\text{warmup}} \leftarrow \text{LoRA}(f_{\text{base}}, D_{\text{warmup}})$
  // Step 2: Gradient Computation
  for each training data $z_i \in D$ **do**
    $g_i \leftarrow \nabla_{\theta_w} L(f_{\text{warmup}}(z_i), y_i)$
    $\tilde{g}_i \leftarrow \text{Normalize}(Rg_i)$ {Random projection}
  end for
  for each validation data $z'_j \in D^k_{\text{val}}$ **do**
    $g'_j \leftarrow \nabla_{\theta_w} L(f_{\text{warmup}}(z'_j), y'_j)$
    $\tilde{g}'_j \leftarrow \text{Normalize}(Rg'_j)$
  end for
  // Step 3: Influence Matrix Computation
  for each $z_i \in D$, $z'_j \in D^k_{\text{val}}$ **do**
    $I^k_{ij} \leftarrow \langle \tilde{g}_i, \tilde{g}'_j \rangle$
  end for
  // Compute average influence per training sample
  $\bar{I}_k(z_i) \leftarrow \frac{1}{|D^k_{\text{val}}|} \sum_{j=1}^{|D^k_{\text{val}}|} I^k_{ij}$
end for
return Task-specific influence scores $\{\bar{I}_k\}_{k=1}^{K}$

Algorithm 2 ICONS Stage 2: Generalist (Influence Consensus-based Data Selection)

Require: Task-specific influence scores $\{\bar{I}_k\}_{k=1}^{K}$
Require: Selection ratio p, number of tasks K
Ensure: Selected subset $S \subset D$ of size $m \ll N$

// Compute voting thresholds
for each task $T_k \in T$ **do**
  $\tau_k \leftarrow (1-p)$-th percentile of $\{\bar{I}_k(z_i)\}_{i=1}^{N}$
end for
// Voting process
for each training sample $z_i \in D$ **do**
  $I_{\text{vote}}(z_i) \leftarrow 0$
  for each task $T_k \in T$ **do**
    $\text{vote}_k(z_i) \leftarrow \mathbb{1}[\bar{I}_k(z_i) \geq \tau_k]$
    $I_{\text{vote}}(z_i) \leftarrow I_{\text{vote}}(z_i) + \text{vote}_k(z_i)$
  end for
end for
// Select top samples based on total votes
$S \leftarrow$ top-p samples by I_{vote}
return Selected subset S

D. Future Work

Our work opens several promising research directions for improving vision-language data selection. While our work focuses specifically on visual instruction tuning data, our influence consensus approach can be naturally extended to other stages of MLLM training, such as alignment stage. The majority voting mechanism may under-represent tasks with unique characteristics or those in the long tail, as it prioritizes samples that broadly benefit multiple tasks to build the *main knowledge pool*. This can lead to limited support for specialized tasks or the reinforcement of spurious correlations spanning multiple tasks. Future work could explore **weighted voting mechanisms**, in which tasks are assigned weights based on their relative importance or contribution to overall model performance for more balanced data selection. Additionally, investigating more efficient gradient computation and storage methods would help scale these methods to larger datasets while maintaining strong performance across diverse vision-language tasks.

E. Visualizations

[Xindi: fix3]

We visualize the most influential top 3 examples across specialists and the generalist selection. The visualization of top influential examples reveals distinct patterns in what makes training samples valuable for different vision-language tasks. VQAv2, GQA, and SQA-I specialists favor multi-turn Q&A scenarios that test both visual comprehension and contextual understanding, while TextVQA, POPE, and MME specialists emphasize text recognition, object verification, and spatial relationships respectively. MMBench-EN and MMBench-CN show consistent patterns despite language differences, focusing on clear, unambiguous scenes that translate well. The LLaVA-W Bench specialist prioritizes examples requiring detailed explanations and multi-step reasoning, and the answers are generally longer. The generalist model values diverse scenarios that combine multiple skills simultaneously. Common characteristics that make these examples particularly valuable include multi-turn interactions, clear visual elements, factual and inferential reasoning, cross-modal interaction, and the ability to test multiple capabilities within a single example. This suggests that the most effective training samples are those that combine multiple types of reasoning while maintaining clear, unambiguous ground truth that can be consistently learned across different tasks. We hope those insights can help guide the future instruction tuning data collection.

Specialist: VQAv2 Selected Top 1 Example

Image:

Conversations:

- <image>\n What does the sofa appear to be blocking?
  Window
- What is the object in the upper right?
  Lamp
- What is the cat sitting on?
  Couch
- What animal is on the couch?
  Cat
- Is the cat sleeping?
  Yes
- Do the cushions match the color of the couch?
  No
- What is the color of the side lamps?
  Pink
- Does the couch have legs under it?
  No
- What is on the couch?
  Cat
- Is there a bed?
  No
- What's on the sofa?
  Cat
- Are the cushions golden?
  No
- How many pillows are on the couch?
  2
- What kind of store is the person in?
  None
- Who sleeps on the bed?
  Cat
- How many pillows in the shot?
  2
- What color is the pillow on the right?
  Blue
- What is covering the couch?
  Cat

Specialist: VQAv2 Selected Top 2 Example

Image:

Conversations:

- <image>\n Is this toaster made out of stainless steel?
  Yes
- Why would someone leave this toaster on the sidewalk?
  Broken
- Is the street cracked?
  Yes

Specialist: VQAv2 Selected Top 3 Example

Image:

Conversations:

- <image>\n Which national flag is in the segment screen of this broadcast? A. netherlands B. german C. france D. uk Answer with the option's letter from the given choices directly.
  D

Specialist: GQA Selected Top 1 Example

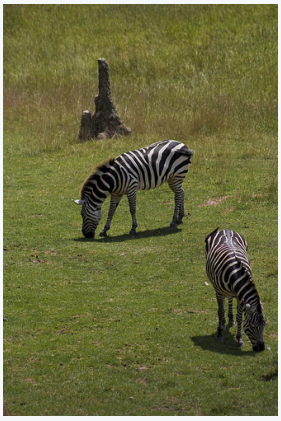
Image:

Conversations:

- Which place is it?
  Airport
- On which side is the man?
  Left

Specialist: GQA Selected Top 2 Example

Image:



Conversations:

- Please provide the bounding box coordinate of the region this sentence describes: a large rock in back.
  $[0.31, 0.08, 0.49, 0.35]$
- Please provide the bounding box coordinate of the region this sentence describes: black stripes on zebra fur.
  $[0.43, 0.37, 0.58, 0.42]$
- Please provide a short description for this region: $[0.33, 0.32, 0.63, 0.58]$.
  A small zebra with big black stripes.
- Please provide a short description for this region: $[0.33, 0.11, 0.48, 0.34]$.
  Rock structure in grass.
- Please provide a short description for this region: $[0.67, 0.76, 0.76, 0.84]$.
  Black zebra hooves.
- Please provide the bounding box coordinate of the region this sentence describes: the zebra is eating grass.
  $[0.31, 0.44, 0.42, 0.58]$
- Please provide the bounding box coordinate of the region this sentence describes: a zebra having a snack.
  $[0.65, 0.54, 0.81, 0.85]$
- Please provide a short description for this region: $[0.61, 0.77, 0.79, 0.86]$.
  A zebra shadow in grass.
- Please provide a short description for this region: $[0.31, 0.32, 0.66, 0.58]$.
  An animal standing in place.
- Please provide the bounding box coordinate of the region this sentence describes: a striped zebra eating.
  $[0.33, 0.32, 0.63, 0.57]$

Specialist: GQA Selected Top 3 Example

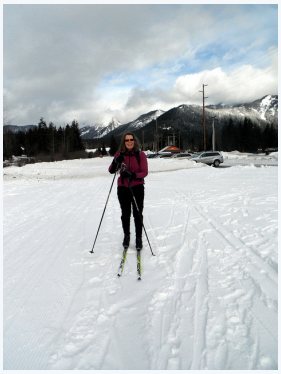
Image:

Conversations:

- Who is the catcher in front of?
  Men
- Who is in front of the men that are wearing a helmet?
  Catcher
- Who is wearing a uniform?
  Men
- Are the men to the left of the catcher wearing a cap?
  No
- Who is wearing a helmet?
  Men
- Who wears the face mask?
  Men
- Who is wearing helmets?
  Catcher
- What kind of clothing is black?
  Uniform
- Are the trousers white?
  Yes
- Who wears a face mask?
  Men
- Do the number and the shirt have the same color?
  Yes
- Which place is it?
  Stadium
- What place was the image taken at?
  Stadium
- What is the person in front of the men wearing?
  Helmets
- Does that uniform look blue?
  No
- What is the catcher wearing?
  Helmets
- Which color is the shirt?
  White
- What is the black clothing item called?
  Uniform
- Is the color of the uniform different than the belt?
  No
- Who is wearing the uniform?
  Men

Specialist: VizWiz Selected Top 1 Example

Image:



Conversations:

- `<image>\n` Is this a man?
  No
- Is there a parking lot?
  Yes
- Does this person look upset?
  No
- Is the lady going up or down the mountain?
  Down

Specialist: VizWiz Selected Top 3 Example

Image:

Conversations:

- `<image>\n` What are the people holding in their hands?
  Food
- Is this a teddy bear?
  No
- Is the girl feeding the elephant?
  No
- Why are these people gathered together?
  Zoo
- Is this some kind of celebration?
  No
- Are there flowers on the girl's dress?
  Yes
- Is this someone's home?
  No
- Where are the people in the picture?
  Zoo
- Was this photo taken in a zoo?
  Yes
- Is the sky clear?
  Yes
- Is it summertime?
  Yes

Specialist: VizWiz Selected Top 2 Example

Image:

Conversations:

- `<image>\n` Would this be a good entree?
  No
- What is the white cloth that those snacks are sitting on?
  Doily
- Would this be an appropriate food to serve to an infant?
  No

Specialist: SQA-I Selected Top 1 Example

Image:

Conversations:

- <image>\n Are there pipes in the bathroom?
  Yes
- What kind of floor is in the photo?
  Black
- What is color of the toilet seat?
  White
- Is the toilet functional?
  Yes

Specialist: SQA-I Selected Top 3 Example

Image:

Conversations:

- How many stools are next to the kitchen counter or island? \n<image>
  There are two stools next to the kitchen counter or island.
- What color is the refrigerator in the kitchen?
  The refrigerator in the kitchen is white.
- What type of material is used for the table and chairs?
  The table and chairs are made of wood.
- Can you describe the layout of the living space and kitchen?
  The living space and kitchen have an open floor plan, which means they are situated next to each other without any wall separations or partitions. This design creates a spacious, airy, and connected environment that allows for easy socialization and movement between the two spaces. In the kitchen, there is a small island with two stools near it, providing a casual seating area for dining or conversation. The wooden table and chairs, along with the white refrigerator, further contribute to the overall aesthetic and layout of the combined living and kitchen area.

Specialist: SQA-I Selected Top 2 Example

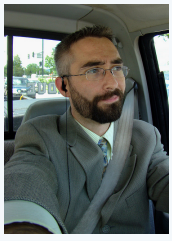
Image:

Conversations:

- <image>\n What is Santa using?
  Laptop
- What is the Santa doing?
  Typing
- Could it be Christmas?
  Yes
- Is this indoors?
  Yes

Specialist: TextVQA Selected Top 1 Example

Image:



Conversations:

- <image>\n The man wearing the suit and tie is operating what object? A. sedan B. coupe C. pickup truck D. suv Answer with the option's letter from the given choices directly.
  C

Specialist: TextVQA Selected Top 2 Example

Image:

Conversations:

- <image>\n Please provide a short description for this region: [0.21, 0.4, 0.41, 0.74].
  Bird biggest.

Specialist: TextVQA Selected Top 3 Example

Image:

Conversations:

- <image>\n Please provide the bounding box coordinate of the region this sentence describes: man in a brown tshirt.
  [0.21, 0.31, 0.34, 0.76]
- Please provide a short description for this region: [0.31, 0.32, 0.48, 0.73].
  This is a man wearing a blue shirt, reaching to shake hands with his team mate.
- Please provide the bounding box coordinate of the region this sentence describes: man standing with raquet in white shirt on near side of net.
  [0.66, 0.36, 0.81, 0.78]
- Please provide the bounding box coordinate of the region this sentence describes: smiling black dude congratulates the winners of a tennis game.
  [0.31, 0.32, 0.48, 0.73]
- Please provide a short description for this region: [0.66, 0.36, 0.81, 0.78].
  A man with a white outfit and a tennis racket with blue handle.
- Please provide a short description for this region: [0.21, 0.31, 0.34, 0.76].
  A man in a brown shirt.
- Please provide the bounding box coordinate of the region this sentence describes: a man in a white shirt holding his hand out.
  [0.53, 0.35, 0.71, 0.73]
- Please provide the bounding box coordinate of the region this sentence describes: the bald older white man playing tennis.
  [0.53, 0.35, 0.71, 0.73]

Specialist: POPE Selected Top 1 Example

Image:

Conversations:

- <image>\n What shape is the dogs tag?<image>\nAnswer the question using a single word or phrase.
  Bone
- What kind of dog is this?
  Beagle
- What is the dog looking in?
  Mirror

Specialist: POPE Selected Top 2 Example

Image:

Conversations:

- <image>\n Is the woman happy? \nAnswer the question using a single word or phrase.
  Yes
- What sport is this girl trying out?
  Snowboarding
- Does the girl have a serious face?
  No
- Is she cold?
  Yes
- Is there snow everywhere?
  No
- What color are the women's gloves?
  White
- What is on the girl's face?
  Smile
- Is this man made snow?
  Yes

Specialist: POPE Selected Top 3 Example

Image:

Conversations:

- <image>\n On which side of the picture is the white dog?\nAnswer the question using a single word or phrase.
  Left
- Do you see any dogs or cats there?
  Yes
- What is under the counter top?
  Drawers
- What is under the countertop?
  Drawers
- What kind of furniture is under the counter top?
  Drawers
- What kind of furniture is under the countertop?
  Drawers
- Is there a white dog or cat?
  Yes
- Are there both cabinets and bowls in the picture?
  No
- Does the trash can look large and white?
  No
- How large is the trashcan on the left side?
  Small
- Does the garbage can have a different color than the dog?
  No
- What animal has the same color as the small trash can?
  Dog
- Are there any ovens or bowls in the photo?
  No
- Which side of the image are the drawers on, the right or the left?
  Right
- What animal is it?
  Dog

Specialist: MME Selected Top 1 Example

Image:

Conversations:

- <image>\n Does the small bike look orange?\n Answer the question using a single word or phrase.
  Yes
- Is the color of the seat different than the sticker?
  Yes

Specialist: MME Selected Top 2 Example

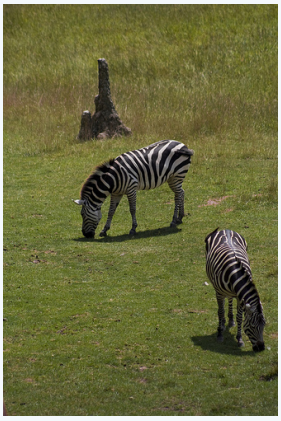
Image:

Conversations:

- <image>\n Which color does the sky have?\nAnswer the question using a single word or phrase.
  Blue
- What is the tree covered in?
  Leaves

Specialist: MME Selected Top 3 Example

Image:

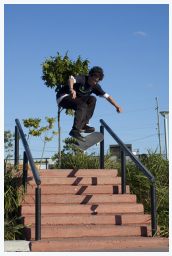


Conversations:

- Please provide the bounding box coordinate of the region this sentence describes: a large rock in back.
  $[0.31, 0.08, 0.49, 0.35]$
- Please provide the bounding box coordinate of the region this sentence describes: black stripes on zebra fur.
  $[0.43, 0.37, 0.58, 0.42]$
- Please provide a short description for this region: $[0.33, 0.32, 0.63, 0.58]$.
  A small zebra with big black stripes.
- Please provide a short description for this region: $[0.33, 0.11, 0.48, 0.34]$.
  Rock structure in grass.
- Please provide a short description for this region: $[0.67, 0.76, 0.76, 0.84]$.
  Black zebra hooves.
- Please provide the bounding box coordinate of the region this sentence describes: the zebra is eating grass.
  $[0.31, 0.44, 0.42, 0.58]$
- Please provide the bounding box coordinate of the region this sentence describes: a zebra having a snack.
  $[0.65, 0.54, 0.81, 0.85]$
- Please provide a short description for this region: $[0.61, 0.77, 0.79, 0.86]$.
  A zebra shadow in grass.
- Please provide a short description for this region: $[0.31, 0.32, 0.66, 0.58]$.
  An animal standing in place.
- Please provide the bounding box coordinate of the region this sentence describes: a striped zebra eating.
  $[0.33, 0.32, 0.63, 0.57]$

Specialist: MMBench Selected Top 1 Example

Image:



Conversations:

- `<image>\n` Why is the man midair in the middle of the steps? A. was thrown B. he fell C. stood up D. performing trick Answer with the option's letter from the given choices directly.
  D

Specialist: MMBench Selected Top 2 Example

Image:

Conversations:

- `<image>\n` Is the machine turned on?\nAnswer the question using a single word or phrase.
  Yes
- Where is this machine used?
  Kitchen
- Can you see through the bottle on the right?
  Yes

Specialist: MMBench Selected Top 3 Example

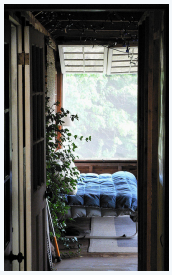
Image:

Conversations:

- <image>\n What type of bread is this?\nAnswer the question using a single word or phrase.
  Banana
- What fruit is in the back?
  Bananas
- Is the fruit ripe?
  Yes

Specialist: MMBench-CN Selected Top 2 Example

Image:



Conversations:

- <image>\n Is the window open?\nAnswer the question using a single word or phrase.
  Yes
- Could the bed's metal leg be broken?
  No
- What color are the blinds/shutters?
  White
- What is seen in the room?
  Bed
- What kind of tree is closest to the house?
  Oak
- What is the blue object?
  Bed

Specialist: MMBench-CN Selected Top 1 Example

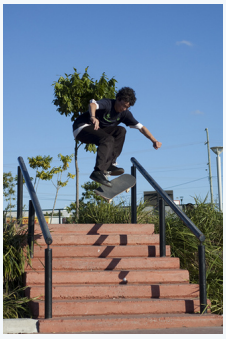
Image:

Conversations:

- <image>\n Is the machine turned on?\nAnswer the question using a single word or phrase.
  Yes
- Where is this machine used?
  Kitchen
- Can you see through the bottle on the right?
  Yes

Specialist: MMBench-CN Selected Top 3 Example

Image:



Conversations:

- <image>\n Why is the man midair in the middle of the steps? A. was thrown B. he fell C. stood up D. performing trick Answer with the option's letter from the given choices directly.
  D

Specialist: LLaVA-W Bench Selected Top 1 Example

Image:

Conversations:

- `<image>\n` Who is in the image?
  In the image, there is a person wearing skis and a dog, both out in the snow.
- Based on the image, what activity is the person likely about to engage in?
  Based on the image, the person is likely about to engage in skiing with their dog.
- What is the context of the photo? Where is it likely taken?
  The context of the photo is that it is being taken outside a ski patrol building. The skier and their dog are likely in a ski resort or mountainous area suitable for skiing.
- Could the skier be part of the ski patrol team?
  It is possible that the skier could be part of the ski patrol team, especially since they are posing outside of a ski patrol building. However, there isn't enough information in the image to say for certain whether they are a ski patrol official.
- How could a dog participate or contribute to skiing activities?
  Dogs can participate in skiing activities in several ways, often as companions or support animals. Some common ways dogs join skiing activities are:
  - Skijoring: In this winter sport, a person on skis is pulled by a dog (or a team of dogs) on a harness. Skijoring combines aspects of cross-country skiing and dog sledding, and both the dog and the skier work together to navigate snowy trails.
  - Ski patrol dogs: Ski patrol teams may have dogs trained for search and rescue missions, avalanche response, and providing comfort to injured skiers. These dogs are typically of large breeds like German Shepherds, Golden Retrievers, or Labrador Retrievers, and are specifically trained to handle challenging winter conditions.
  - Companion dogs: Many dog owners bring their dogs along for recreational cross-country or backcountry skiing. These dogs do not necessarily have formal training but are physically capable of handling cold temperatures and snowy terrain, and they accompany their owners as companions on the trails.

  In the image provided, it's unclear what specific role the dog may play or how they will participate in the skiing activity. However, it's evident that the dog will accompany the skier during their skiing session.

Specialist: LLaVA-W Bench Selected Top 2 Example

Image:

Conversations:

- `<image>\n` Why might passengers need to pay attention to the surroundings at the train station?
  At the train station, passengers should be alert to their surroundings for various reasons. In the image, a train is moving along the tracks, and a green light indicates that it is safe to proceed. However, there are several elements to be aware of in the scene, such as other people near the tracks, cars nearby, and potential transit-related changes signaled by the traffic lights. Being attentive to the surroundings can help people anticipate changes in train schedules, ensure they board the correct train, and avoid accidents or other safety hazards. Additionally, it is essential to be mindful of fellow travelers' presence and maintain a safe distance from the tracks to ensure both personal safety and the well-being of others. This can help create a safe and efficient environment at the train station for everyone.

Specialist: LLaVA-W Bench Selected Top 3 Example

Image:

Multi-turn Conversations:

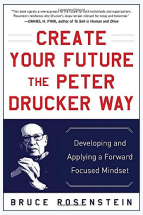
- What are the key elements in this picture?\n<image>
  The image features a city street where a long row of parked motorcycles and mopeds are lined up along the sidewalk. There are several motorcycles of various sizes and designs, creating a neat row that extends across the side of the street.

  There is also a notable presence of pedestrians in the area, as numerous people can be seen walking around and engaging in different activities. A backpack is located close to one of the people, placed on the pavement.

  In addition to the motorcycles, other vehicles are present on the street, including a car and a truck parked near the sidewalk, adding to the urban scene's atmosphere.

Generalist Selected Top 1 Example

Image:



Conversations:

- <image>\n Who is the author of this book?
  Bruce Rosenstein
- What is the title of this book?
  Create Your Future the Peter Drucker Way: Developing and Applying a Forward-Focused Mindset
- What type of book is this?
  Business & Money
- Is this a financial book?
  Yes
- Is this a kids book?
  No

Generalist Selected Top 2 Example

Image:

Conversations:

- <image>\n What could be a possible use case for the items on the desk?
  A possible use case for the items on the desk is a home office or a gaming setup. The desk has a computer monitor, keyboard, mouse, and a headset, which indicates that it is a workstation that could be used for tasks requiring a personal computer. These tasks could involve remote work, video editing, software development, or other professional activities. In addition, the presence of the headset and the gaming peripherals such as the mouse and keyboard can also suggest a gaming setup, where the person uses their computer for playing video games, communicating with other players using the headset, or even streaming their gameplay online. The desk might also be used for general computer use, such as browsing the internet, watching movies, or using social media platforms, as evidenced by the reference to Flickr on the computer monitor.

Generalist Selected Top 3 Example

Image:

Conversations:

- `<image>\n` What precautions should these young adults take to ensure safety while playing the Wii game?
  To ensure safety while playing the Wii game, these young adults should maintain an adequate distance from each other, the TV, and other objects in the room. With the use of the Wii remotes, they should have sufficient space for movement to prevent accidentally hitting or running into each other, furniture, or nearby items. Furthermore, they should consider using wrist straps for the remotes to reduce the risk of accidentally throwing the controller during gameplay. Finally, keeping the play area clean and free from obstructions, such as loose cables or objects on the floor, can help minimize the risk of trip hazards and improve safety.

F. Cambrian Experiments

[Xindi: will move this section to main after we have the complete results]

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-W Bench	Rel. (%)
Full (7068k)	80.2	62.9	58.4	75.3	60.9	86.5	1,524.6	69.1	58.9	67.6	
Random ICONS (ours)											

Table 9. **Cambrian-1.4M Results on LLaVA v1.5.** [Xindi: Can we only include FULL, RANDOM, AND OURS? Since Random is already one of the strongest baseline.]