Does the Question Really Matter? Training-Free Data Selection for Vision-Language SFT

Peng Sun[1] **Yi Yang**[1] **Huawen Shen**[2] **Yi Ban**[1]
Yanbo Wang[3] **Yuqiang Li**[4] **Tianfan Fu**[1]*
[1]State Key Laboratory for Novel Software Technology,
School of Computer Science, Nanjing University, Nanjing, China
[2]Institute of Information Engineering,
Chinese Academy of Sciences, Beijing, China
[3]Institute of Big Data and Visual Computing,
North University of China, Taiyuan, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract

Visual instruction tuning is crucial for improving vision-language large models (VLLMs). However, many samples can be solved via linguistic patterns or common-sense shortcuts, without genuine cross-modal reasoning, limiting the effectiveness of multimodal learning. Prior data selection methods often rely on costly proxy model training and focus on difficulty or diversity, failing to capture a sample's true contribution to vision-language joint reasoning. In this paper, we propose CVS, a training-free data selection method based on the insight that, for high-quality multimodal samples, introducing the question should substantially alter the model's assessment of answer validity given an image. CVS leverages a frozen VLLM as an evaluator and measures the discrepancy in answer validity with and without conditioning on the question, enabling the identification of samples that require vision-language joint reasoning while filtering semantic-conflict noise. Experiments on Vision-Flan and The Cauldron show that CVS achieves solid performance across datasets. On Vision-Flan, CVS outperforms full-data training by 3.5% and 4.8% using only 10% and 15% of the data, respectively, and remains robust on the highly heterogeneous Cauldron dataset. Moreover, CVS reduces computational cost by 17.3% and 44.4% compared to COINCIDE and XMAS.

1 Introduction

One of the key sources of the capabilities of vision-language large models (VLLMs) is Visual Instruction Tuning (VIT) [22], which enables models to better align natural language instructions with visual content across a wide range of tasks [20, 53]. However, a gradually emerging yet underexplored issue is that, during visual instruction tuning, not all samples that are formally multimodal actually require genuine cross-modal reasoning [45].

Among existing datasets, many samples can be answered correctly with little or no reliance on visual content, as models exploit linguistic patterns or common-sense priors [28, 39]. Such samples provide weak cross-modal supervision and encourage models to adopt linguistic shortcuts rather than visual reasoning, ultimately degrading sensitivity to visual evidence [52]. This issue is exacerbated in

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

real-world data construction pipelines, where heterogeneous collection methods introduce supervision noise and image–text misalignment, rendering simple data scaling insufficient for improving vision-language models [6, 49].

Existing data selection methods can be categorized into score-based and clustering-based methods. The former typically assign scalar scores to samples based on criteria such as difficulty, gradient contribution, or similarity, but often treat sample utility as an independently evaluable property, making it difficult to distinguish samples that genuinely require vision-language joint reasoning from those that can be solved via linguistic shortcuts [19]. The latter aim to improve data diversity through clustering or deduplication; however, diversity alone does not guarantee that a question meaningfully constrains its answer [42]. Moreover, most existing methods rely on additional proxy models or complex post-processing pipelines, which introduce substantial computational overhead in large-scale data settings [2]. We posit that a sample's value implies Visual Necessity: the question must provide non-redundant information gain regarding the answer's validity, conditioned on the visual context. Only when the question substantively affects the decision process and forces the model to jointly reason over visual content and linguistic instructions does the supervision signal become effective. This conditional dependence captures the essence of cross-modal supervision, yet has not been explicitly modeled in existing data selection methods.

Recent advances in VLLMs show that modern models are already capable of making stable zero-shot judgments about answer validity given an image and text [9, 35]. Motivated by this observation, we adopt a reverse perspective. Instead of training an additional proxy model to assess sample quality, we directly leverage a frozen VLLM as an evaluator and use changes in its judgment behavior to characterize the supervisory value of a sample. Based on this insight, we propose Conditional Verdict Shift (CVS), a training-free data selection method. CVS evaluates sample value by comparing the model's judgments of answer validity under two conditions. One condition uses the full context, including the image, question, and answer, while the other removes the question and retains only the image and the answer. If the inclusion of the question increases the model's acceptance of the answer without strengthening its tendency to reject it, the sample is regarded as semantically consistent and provides effective cross-modal constraints. Otherwise, it is more likely to reflect semantic conflicts or noisy supervision. Through this conditional verdict shift, CVS filters low-quality samples and prioritizes high-value ones that require genuine cross-modal reasoning, without performing any additional training.

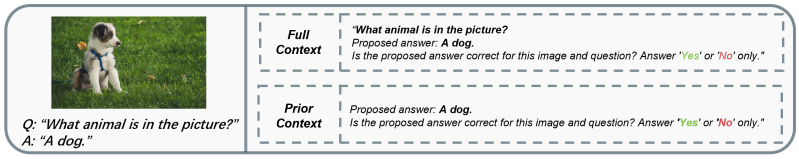
Our main contributions are summarized as:

- We identify a critical yet often overlooked issue in visual instruction data: a large number of samples do not genuinely rely on vision-language joint reasoning, but can instead be solved via linguistic shortcuts, thereby weakening the effectiveness of cross-modal learning.
- We propose Conditional Verdict Shift (CVS), a training-free data selection method based on a novel perspective that models the effectiveness of VIT data through the conditional influence of the question on answer validity. This allows efficient identification of high-value samples requiring genuine joint reasoning.
- We systematically evaluate CVS on two widely recognized VIT datasets, demonstrating its advantages in performance, stability, and computational efficiency. On VISION-FLAN, using only 10% and 15% of samples selected by CVS yields 3.5% and 4.8% higher performance than full-data training, respectively. Furthermore, on The Cauldron, CVS reduces computational costs by 17.3% and 44.4% compared to COINCIDE and XMAS.

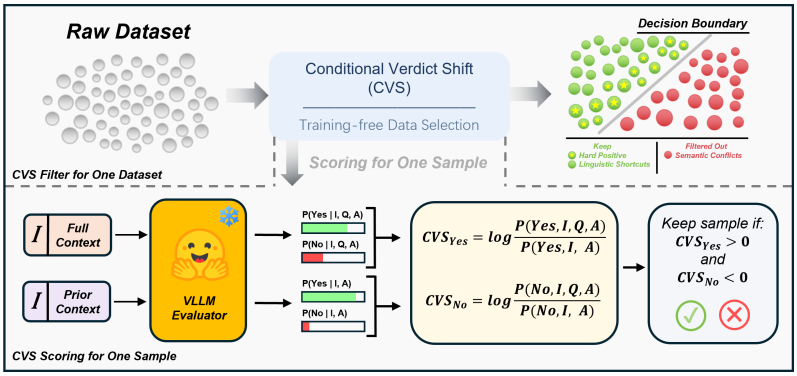
2 Method

2.1 Problem formulation

Appendix B analyzes data selection from a distributional perspective and shows that, when a sufficiently strong evaluator is available, satisfactory generalization can be achieved using only a small



(a) Prompt construction.



(b) CVS pipeline.

Figure 1: Overview of CVS. (a) Prompt construction. (b) CVS uses a frozen vision-language model to estimate conditional verdict shifts, filter noisy samples, and prioritize informative samples near the decision boundary.

selected subset. This section does not introduce additional theoretical objectives or optimization assumptions. Rather, it reformulates the data selection problem from an operational discriminative perspective, thereby clarifying the core question that motivates the subsequent CVS design.

We consider a candidate training pool

$$\mathcal{S}_{\text{pool}} = \{s_i\}_{i=1}^{N}, \quad s_i = (I_i, Q_i, A_i), \tag{1}$$

where I_i, Q_i, and A_i denote the image, question, and answer of the i-th sample, respectively. Since candidate samples may contain heterogeneous supervision signals, such as linguistic shortcuts, weak visual grounding, or semantic conflicts, we do not explicitly model the target data distribution. Instead, we introduce a binary selection variable

$$m_i \in \{0, 1\}, \tag{2}$$

where $m_i = 1$ indicates that s_i is retained for training. The selected training subset is therefore

$$\mathcal{S}_{\text{train}} = \{s_i \in \mathcal{S}_{\text{pool}} \mid m_i = 1\}. \tag{3}$$

To characterize the desired selection criterion, we define an abstract sample value function $\mathcal{V}(s_i)$, which measures the extent to which a sample contributes useful vision-language learning signals. Under a budget constraint K, data selection can be conceptually written as

$$\max_{\{m_i\}_{i=1}^{N}} \sum_{i=1}^{N} m_i \mathcal{V}(s_i), \quad \text{s.t.} \sum_{i=1}^{N} m_i \leq K, \; m_i \in \{0, 1\}. \tag{4}$$

This formulation is not intended as an optimization objective to be solved directly. Rather, it specifies the type of sample value that an effective selection method should approximate. In this work, the key question is how to instantiate $\mathcal{V}(\cdot)$ in a computable and training-free manner, while keeping it closely aligned with vision-language joint reasoning. The proposed Conditional Verdict Shift (CVS), introduced in Section 2.2, addresses this question by using a frozen vision-language model to measure discriminative changes under different conditioning contexts.

2.2 Conditional verdict shift (CVS)

In our implementation, we use a frozen, well-trained VLLM as an intrinsic evaluator that evaluates the validity of an answer A given its context. This is formulated as a single-step binary classification task over $\mathcal{Y} = \{\text{YES}, \text{NO}\}$. Rather than relying on absolute likelihoods, we characterize the extent to which the textual query Q modulates the model's judgment regarding the validity of the answer A. This design draws inspiration from established methodology [11] that uses conditional V-information to evaluate the semantic quality of instances via pointwise mutual information estimates.

The overall CVS prompt construction and pipeline are illustrated in Figures 1a and 1b, respectively, with detailed prompt templates provided in Appendix C. For binary predictions, we directly use the model's output probabilities for $\{\text{YES}, \text{NO}\}$ without additional softmax normalization. Since different answer-token choices and prompt templates may affect the resulting model-assigned probabilities, we examine the sensitivity of CVS to these design choices in Appendix F to assess its robustness.

Conditional affirmation shift We measure the modulation of confidence induced by the query Q. Specifically, we compute the probability of the token 'YES' under two conditions: the full context (I, Q, A) and a reduced context where the query Q is removed. We formalize this comparison as the *Conditional Affirmation Shift* (CVS_{YES}), defined as the logarithmic ratio:

$$\text{CVS}_{\text{YES}} = \log \frac{P(\text{YES} \mid I, Q, A)}{P(\text{YES} \mid I, A)}. \tag{5}$$

This metric quantifies the information gain provided by the query Q regarding the validity of the answer. A positive shift, i.e., $\text{CVS}_{\text{YES}} > 0$, indicates that the inclusion of the question reinforces the model's belief in the answer's correctness, thereby signifying a strong semantic alignment between the query Q and the response A.

Conditional rejection shift To characterize the "inverse impact" of the textual query on the answer's validity, we further investigate the relative variation in the model's rejection confidence. Analogously to the affirmation shift, we define the *Conditional Rejection Shift* (CVS_{NO}) as:

$$\text{CVS}_{\text{NO}} = \log \frac{P(\text{NO} \mid I, Q, A)}{P(\text{NO} \mid I, A)}. \tag{6}$$

This metric measures whether the introduction of the query Q reinforces the model's tendency to classify the answer as "invalid" or "incorrect". A positive shift ($\text{CVS}_{\text{NO}} > 0$) implies that the query context elevates the probability of rejection, signaling a semantic conflict or misalignment between the question and the answer (e.g., hallucinations or irrelevant responses).

Filtering protocol Combining gains from both affirmation and rejection, we establish a principled data filtering protocol. Specifically, we construct an admissible dataset $\mathcal{D}_{\text{select}}$ by retaining samples that satisfy the following alignment constraints:

$$\mathcal{C}_{\text{align}}: \quad (\text{CVS}_{\text{YES}} > 0) \wedge (\text{CVS}_{\text{NO}} < 0). \tag{7}$$

This protocol enforces semantic consistency by ensuring answer validity while suppressing rejection, filtering inconsistent or misleading QA pairs. We keep the affirmation and rejection shifts as two separate constraints because they capture answer support and semantic conflict from different directions. To examine whether this separation is necessary in practice, we further compare variants using only CVS_{YES}, only CVS_{NO}, and three coupled log-odds variants, with results reported in Appendix D.

For CVS_{YES} and CVS_{NO}, instead of using more specialized thresholding schemes, we adopt zero as a natural and effective threshold, since it directly separates positive from negative shifts. With this choice, we retain 59.29% and 63.97% of samples on the two datasets used in subsequent experiments, respectively.

Preference for hard positive samples Building on our filter protocol, and contrary to intuition, we prioritize samples with lower positive CVS_{YES} rather than maximizing this metric. A high CVS_{YES} suggests that the model can resolve the reasoning path $Q \to I \to A$ with little effort, often due to strong linguistic correlations or commonsense shortcuts that allow it to ignore the visual input I and rely on language priors alone. In contrast, we focus on instances that require joint reasoning, where I provides essential contextual grounding and Q specifies the directive information. A lower CVS_{YES} indicates that the model validates the answer without excessive confidence, reflecting a non-trivial inference process. These samples lie closer to the decision boundary and, during training, force explicit integration of visual features and textual semantics, yielding substantially stronger gradient learning signals than high-confidence samples.

3 Experiments

In this section, we describe the experimental setup in Section 3.1 and analyze the main results in Section 3.2.

3.1 Experimental setup

Visual instruction tuning datasets We conduct experiments on two widely used ViT datasets, Vision-Flan [48] and The Cauldron [17] (see Appendix H for details).

Models and training details We use the pre-trained LLaVA-1.5-7B model [23] as the target model and Qwen2.5-VL-7B-Instruct model [4] as the evaluator, with inference accelerated by vLLM [16]. We choose LLaVA-1.5-7B because its established training protocol supports reproducibility and comparison with prior visual instruction data selection methods. We choose Qwen2.5-VL-7B-Instruct for its open weights and access to the output probabilities required by CVS. Using a different model family for the evaluator may also reduce biases shared with the target model. Unless otherwise specified (Section 4.3), we follow the official LLaVA-1.5 LoRA fine-tuning hyperparameters. All experiments are conducted on 4 NVIDIA RTX A6000 GPUs.

Baselines In the main experiments, we compare CVS with CLIP-Score [13], EL2N [38], SemDeDup [1], D2 Pruning [30], Random, COINCIDE [18], and XMAS [36]. Notably, COINCIDE (2024) and XMAS (2025) achieved state-of-the-art performance at the time of their introduction. Detailed hyperparameter settings for these baselines are provided in Appendix I. We also include two recent data selection methods, SCALSELECT [44] and PRISM [5], as additional baselines. Since these additional baselines do not change the main conclusion and exhibit relatively weak performance, we defer their detailed results to Appendix L. Analysis and ablation studies primarily compare CVS with random sampling (see Appendix J for the rationale).

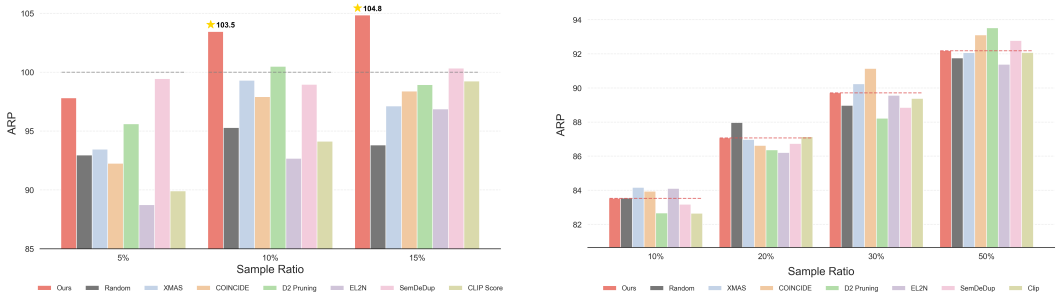
Evaluation We evaluate fine-tuned models on 12 benchmarks spanning general visual question answering, text-intensive visual understanding, document and chart understanding, and multimodal reasoning, including GQA [14], VizWiz [12], TextVQA [40], ScienceQA-IMG [27], MME [21], MMBench [24], AI2D [15], ChartQA [32], DocVQA [33], InfoVQA [34], MMStar [7], and OCR-Bench [25]. Following COINCIDE and XMAS, we report Average Relative Performance (ARP) to normalize evaluation scales across datasets:

$$\text{ARP} = \frac{\text{Subset Data Performance}}{\text{Full Data Performance}} \times 100\%. \tag{8}$$

Detailed descriptions of all benchmarks and complete experimental results are provided in Appendix K and Appendix L, respectively.

3.2 Main results

Vision-Flan Due to its 187 diverse tasks, Vision-Flan poses substantial challenges for multi-task data selection. As shown in Figure 2a, at a 5% sampling ratio, CVS achieves the second-best



(a) Vision-Flan. (b) The Cauldron.

Figure 2: Performance comparison between CVS and baseline methods on Vision-Flan and The Cauldron. CVS performs best at 10% and 15% sampling ratios on Vision-Flan, even surpassing full-data training, and remains robust across sampling ratios on The Cauldron.

performance, likely because its focus on near-boundary samples is less effective before a robust feature space is established under extremely limited data. Once the sampling ratio increases to 10% and 15%, this cold-start issue is mitigated, and the advantage of CVS becomes clear: it achieves state-of-the-art performance, significantly outperforming all baselines. In particular, CVS attains ARP of 103.5 and 104.8 at 10% and 15% sampling, respectively, surpassing even full-data fine-tuning (ARP = 100). These results strongly support our claim in Section 2.2 that filtering out language-prior-dominated samples while prioritizing high-value, decision-boundary samples requiring vision-language joint reasoning leads to more effective gradient signals. Compared to XMAS and COINCIDE, CVS shows more robust performance advantages in the highly multi-task Vision-Flan setting, with the advantage increasing at medium and larger data budgets, indicating improved mitigation of task bias while maintaining broad visual coverage. Another noteworthy phenomenon is that baseline methods (e.g., XMAS and D2 Pruning) show large performance fluctuations as the sampling ratio increases, failing to achieve monotonic gains with larger data budgets. This instability suggests limited ability to distinguish informative samples from noise when expanding the subset. In contrast, CVS exhibits stable and consistent performance improvements across sampling ratios, indicating its robustness in selecting high-quality samples at different data scales. EL2N and CLIP-Score perform poorly across all sampling ratios, consistent with observations [18] that single-metric selection leads to biased and redundant subsets.

The Cauldron Compared with Vision-Flan, The Cauldron exhibits a substantially different form of supervision noise. Specifically, many of its visual instructions are generated via templated or rule-based procedures, leading to pronounced structural redundancy in the data distribution rather than noise dominated by linguistic shortcuts or semantic conflicts. This characteristic shifts the focus of data selection in a qualitatively different way and imposes distinct requirements on methods' inductive biases. As shown in Figure 2b, CVS shows stable and monotonic improvements across sampling ratios, indicating consistent and reliable selection across all budgets on The Cauldron with different noise types. However, at higher sampling ratios (e.g., 30% and 50%), some clustering-based methods (e.g., D2 Pruning and COINCIDE) slightly outperform CVS. We attribute this to the dominance of structural redundancy in The Cauldron: under larger data budgets, coverage and redundancy compression become increasingly influential, favoring clustering-based methods. In contrast, CVS focuses on identifying samples that provide discriminative signals for vision-language joint reasoning by suppressing semantic shortcuts and filtering semantically inconsistent data. As a result, when structural redundancy is the primary noise source, CVS is complementary to clustering- or deduplication-based methods rather than aiming for universal superiority across all budgets.

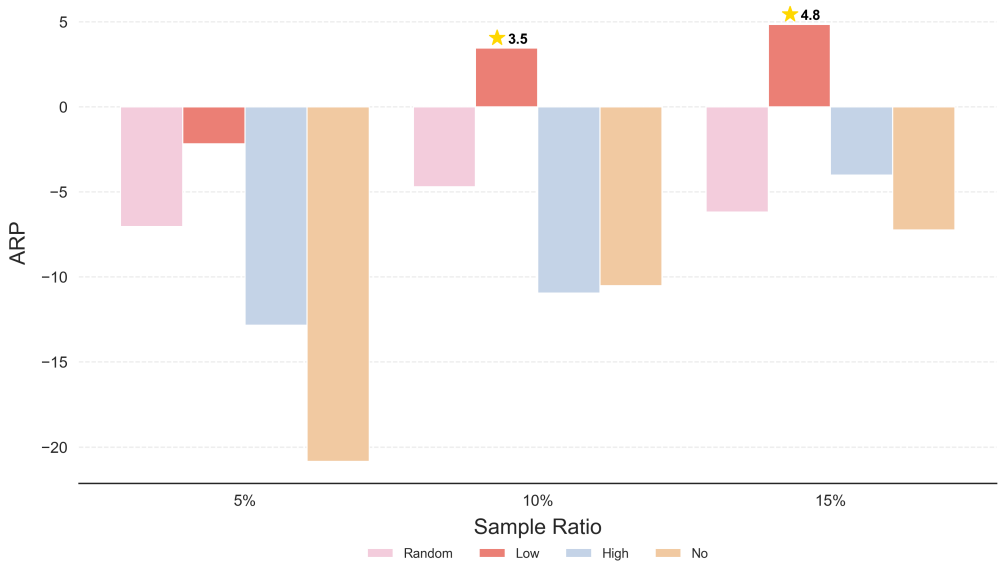


Figure 3: Effect of CVS score ranges on model performance. The low-score strategy consistently yields the largest gains across all budgets.

4 Analysis and ablation studies

4.1 Effect of CVS score ranges

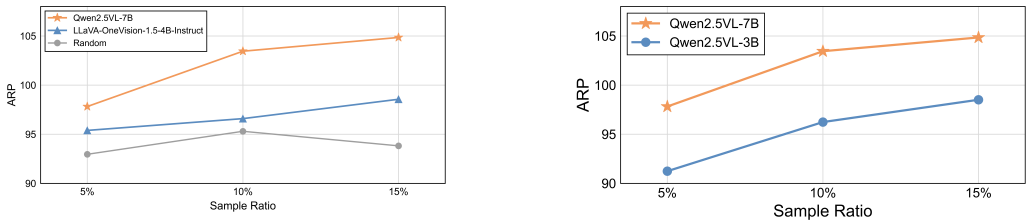
To study the impact of different CVS score ranges on model performance, we conduct a controlled comparison of sampling strategies on Vision-Flan using the pre-trained LLaVA-1.5-7B model. We compare Random sampling against three CVS-based variants: Low selects samples with the lowest CVS_{YES} from the filtered candidate pool, corresponding to samples near the decision boundary; High selects samples with the highest CVS_{YES} from the filtered candidate pool, representing easy high-confidence samples; and No selects samples with the highest CVS_{NO} over the entire dataset, corresponding to queries that strongly increase the rejection probability. The results in Figure 3 show that sampling strategies lead to markedly different performance.

The Low strategy consistently outperforms the others, even under the cold-start setting with only 5% data. As the sampling ratio increases to 10% and 15%, it becomes the only strategy that yields improvements over full-data fine-tuning, achieving gains of +3.5% and +4.8%, respectively. These results support our claim in Section 2.2 that samples near the decision boundary encourage vision-language joint reasoning and provide more informative optimization signals.

In contrast, the High strategy leads to clear performance degradation, performing significantly worse than random sampling at 10% and 15%. This also aligns with our claim in Section 2.2: such samples exhibit strong linguistic correlations or common-sense shortcuts between the query Q and answer A, allowing the model to predict while largely ignoring the image I. Training on these samples provides limited informational gain and wastes computational budget, ultimately harming generalization.

The No strategy yields the worst performance, consistently underperforming random sampling across all ratios, with especially severe degradation at 5% and 10%. High rejection shift typically reflects semantic misalignment or intrinsic noise (e.g., hallucinations), and forcing the model to fit such inconsistent samples disrupts the learned semantic space, leading to negative transfer.

Overall, the Low strategy maximizes data efficiency by identifying correct samples near the decision boundary. Representative Low, High, and No examples are provided in Appendix M.



(a) Evaluator architecture.

(b) Evaluator scale.

Figure 4: Effect of evaluator architecture and scale on CVS performance.

4.2 Robustness to evaluator

This section evaluates the robustness of CVS with respect to evaluator architecture and evaluator scale. All experiments fix the target model to the pre-trained LLaVA-1.5-7B model and are conducted on Vision-Flan.

Evaluator architecture We replace the evaluator from Qwen2.5-VL-7B-Instruct with LLaVA-OneVision-1.5-4B-Instruct [3], which belongs to a different model family. As shown in Figure 4a, CVS consistently outperforms random sampling across all sampling ratios (5%, 10%, and 15%) under both evaluator architectures, indicating that CVS does not rely on any specific model architecture. We further evaluate CVS with InternVL3-8B [54] as the evaluator and report the corresponding results in Appendix L.

Evaluator scale We further vary the evaluator size within the Qwen2.5-VL family, comparing 3B and 7B models. Figure 4b shows a clear scaling trend: larger evaluators consistently lead to better downstream performance across all sampling ratios.

Overall, these results suggest that CVS benefits from stronger evaluators while remaining robust to evaluator architecture. This behavior contrasts with COINCIDE, which reports performance saturation or degradation when increasing proxy model size, indicating that CVS exhibits more favorable scalability. As the community continues to release increasingly powerful VLLMs, CVS can directly leverage stronger evaluators without modifying the algorithm.

4.3 Robustness to target model

To evaluate whether the selected data exhibits cross-model utility, we replace the target model from LLaVA-1.5-7B with the more capable pre-trained Qwen2-VL-2B [43], while fixing the evaluator as Qwen2-VL-7B-Instruct. All experiments are conducted on Vision-Flan. Given the higher capacity of Qwen2-VL-2B, we adopt a moderate learning rate of $1\mathrm{e}{-4}$, while keeping other hyperparameters unchanged. As shown in Table 1, CVS consistently outperforms random sampling across all sampling ratios (5%, 10%, and 15%), even when applied to the stronger Qwen2-VL-2B target model. These results indicate that CVS captures model-agnostic data utility, selecting informative samples that remain effective across different target models. Similar results are observed when using LLaVA-1.5-13B as the target model (see Appendix L).

4.4 Visual anchoring

CVS selects data by comparing $P(Y \mid I, Q, A)$ and $P(Y \mid I, A)$, where $Y \in \{\text{Yes}, \text{No}\}$. To assess the necessity of including image information in the denominator, we conduct an ablation study that removes I from $P(Y \mid I, A)$, yielding a baseline conditioned only on the textual answer A. The experimental setup follows the main experiments and is conducted on Vision-Flan.

As shown in Table 2, data selected by CVS without visual anchoring consistently underperforms random sampling across all sampling ratios, with a performance gap exceeding 10 percentage points

at the 5% sampling rate. This result highlights the critical role of $P(Y \mid I, A)$ as a visual anchoring term in CVS. We analyze its effect from two perspectives.

Visual anchoring establishes a reasoning baseline The standard CVS formulation

$$\frac{P(Y \mid I, Q, A)}{P(Y \mid I, A)} \tag{9}$$

quantifies the information gain contributed by the query Q under a fixed visual context, thereby identifying cases where visual evidence alone is insufficient and vision-language joint reasoning is required.

Semantic collapse under the degenerate formulation When removing I from the denominator, the formulation degenerates to

$$\frac{P(Y \mid I, Q, A)}{P(Y \mid A)}. \tag{10}$$

Under this degenerate setting, the score reflects to the mutual information between the multimodal input (I, Q) and the textual prior of A. Applying Low strategy then favors samples where (I, Q) contributes little to predicting A, resulting in weakly correlated image–text pairs or low-quality cases where the answer is independent of context (e.g., generic or uninformative text). In contrast, in the original CVS formulation, low positive scores have a different interpretation: they indicate that the query improves confidence in the correct answer in a non-trivial manner, requiring feature alignment grounded in visual priors rather than linguistic shortcuts. Consequently, retaining I in the denominator is essential for CVS to distinguish low-value from high-value samples. Only by measuring the guiding effect of language on top of a visual perception baseline can CVS truly select data that promotes vision-language joint reasoning capability.

4.5 Computation time

The computational overhead of data selection methods critically affects their scalability. Table 3 compares the total GPU hours required by CVS with two recent state-of-the-art baselines, COINCIDE and XMAS, on The Cauldron. CVS demonstrates superior computational efficiency, processing the entire dataset in only 10.5 GPU hours, which saves 17.3% and 44.4% of the computation time compared to COINCIDE (12.7h) and XMAS (18.9h), respectively. This advantage mainly stems from the inference-only design of CVS. Unlike COINCIDE and XMAS, which require training a proxy model on the target dataset, CVS uses only forward inference with a frozen model. Eliminating proxy model training substantially reduces computational cost, enabling CVS to scale to multimodal datasets with millions or even hundreds of millions of samples.

5 Conclusion and limitations

Conclusion In this paper, we propose CVS, a training-free data selection method. CVS is based on the observation that high-value multimodal samples should require the question to meaningfully change the model's judgment of answer validity given the image. By comparing a frozen VLLM's verdicts with and without conditioning on the question, CVS filters semantically inconsistent samples and prioritizes informative hard positives that better encourage vision-language joint reasoning. Experiments on Vision-Flan and The Cauldron demonstrate that CVS achieves strong data efficiency, competitive downstream performance, and lower computational cost compared with recent baselines.

Limitations Despite its effectiveness, CVS has several limitations. First, although CVS is robust across different evaluators, its performance may degrade when the frozen evaluator exhibits particularly severe biases or calibration errors. Second, while CVS avoids proxy-model training, it still requires inference over the candidate pool, leaving room for further efficiency improvements. Third, CVS mainly addresses semantic conflicts and language-prior shortcuts. It is not designed to address

Table 1: Robustness of CVS across different target models. On Vision-Flan, we replace LLaVA-1.5-7B with the stronger Qwen2-VL-2B and compare random sampling and CVS at different sampling ratios using Average Relative Performance (ARP; full data = 100, higher is better). CVS outperforms random sampling at all sampling ratios, indicating that the selected data transfers across target models and remains effective with a stronger target model.

Method	5%	10%	15%
Random	98.34	98.64	97.80
CVS	**99.65**	**100.09**	**99.65**

Table 2: Visual anchoring ablation study. On Vision-Flan, we remove image information I from the denominator $P(Y \mid I, A)$ in CVS, yielding CVS without visual anchoring (CVS w/o VA), and compare this variant with random sampling at different sampling ratios using Average Relative Performance (ARP; full data = 100, higher is better). CVS w/o VA performs substantially worse than random sampling at all sampling ratios, indicating that measuring the question-induced conditional verdict shift under the same visual context is crucial for CVS.

Method	5%	10%	15%
Random	**92.96**	**95.30**	**93.62**
CVS	82.44	89.48	88.44

Table 3: Data selection cost on The Cauldron (100K samples) using a single NVIDIA RTX A6000 GPU. Training-Free indicates whether selection avoids proxy model training; GPU Hours measures total selection time (lower is better). CVS requires 10.5 GPU hours, saving 17.3% and 44.4% compared with COINCIDE and XMAS, respectively.

Method	Training-Free	GPU Hours
XMAS	✗	18.9
COINCIDE	✗	12.7
CVS	✓	**10.5**

every source of data inefficiency and may be less effective when structural redundancy is the primary source. Finally, our evaluation is limited to two datasets. Future work could validate CVS on a broader range of datasets and explore its integration with diversity-based methods to further improve data selection under heterogeneous data distributions.

Acknowledgments and Disclosure of Funding

This work was supported by the State Key Laboratory for Novel Software Technology at Nanjing University. The authors declare no competing interests.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023. URL https://arxiv.org/abs/2303.09540.

[2] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL https://arxiv.org/abs/2402.16827.

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[5] Jinhe Bi, Yifan Wang, Danqi Yan, Wenke Huang, Zengjie Jin, Xiaowen Ma, Sikuan Yan, Artur Hecker, Mang Ye, Xun Xiao, et al. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

[6] Chi Chen, Peng Li, Maosong Sun, and Yang Liu. Weakly supervised vision-and-language pre-training with relative representations, 2023. URL https://arxiv.org/abs/2305.15483.

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

[8] 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, 2024. URL https://arxiv.org/abs/2402.12501.

[9] Sufyan Danish, Abolghasem Sadeghi-Niaraki, Samee Ullah Khan, L Minh Dang, Lilia Tightiz, and Hyeonjoon Moon. A comprehensive survey of vision-language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, page 103623, 2025.

[10] Elvis Dohmatob, Mohammad Pezeshki, and Reyhane Askari-Hemmat. Why less is more (sometimes): A theory of data curation, 2025. URL https://arxiv.org/abs/2511.03492.

[11] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:250340652.

[12] 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.

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

[14] 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.

[15] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

[16] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[17] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[18] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[19] Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144, 2025.

[20] Chen Li, Yixiao Ge, Dian Li, and Ying Shan. Vision-language instruction tuning: A review and analysis, 2023. URL `https://arxiv.org/abs/2311.08172`.

[21] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

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

[23] 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.

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[25] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[26] Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: High-value data selection for visual instruction tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3712–3721, 2025.

[27] 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.

[28] Qiwen Lu, Shengbo Chen, and Xiaoke Zhu. Eliminating language bias in visual question answering with potential causality models. *Expert Systems with Applications*, page 128296, 2025.

[29] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Jiayi Ji, Jie Lou, Debing Zhang, and Rongrong Ji. Mllm-selector: Necessity and diversity-driven high-value data selection for enhanced visual instruction tuning, 2025. URL https://arxiv.org/abs/2503.20502.

[30] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning, 2023. URL https://arxiv.org/abs/2310.07931.

[31] 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, 2023. URL https://arxiv.org/abs/2309.04564.

[32] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[33] 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.

[34] 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.

[35] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[36] Nilay Naharas, Dang Nguyen, Nesihan Bulut, Mohammadhossein Bateni, Vahab Mirrokni, and Baharan Mirzasoleiman. Data selection for fine-tuning vision language models via cross modal alignment trajectories, 2025. URL https://arxiv.org/abs/2510.01454.

[37] JeongYeon Nam, Jinbae Im, Wonjae Kim, and Taeho Kil. Extract free dense misalignment from clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6173–6181, 2025.

[38] 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.

[39] Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, 2025.

[40] 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.

[41] 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.

[42] Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. Diversity measurement and subset selection for instruction tuning datasets, 2024. URL https://arxiv.org/abs/2402.02318.

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men,

Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

[44] Changti Wu, Jiahuai Mao, Yuzhuo Miao, Shijie Lian, Bin Yu, Xiaopeng Lin, Cong Huang, Lei Zhang, and Kai Chen. Scalselect: Scalable training-free multimodal data selection for efficient visual instruction tuning. *arXiv preprint arXiv:2602.11636*, 2026.

[45] Chen Henry Wu, Neil Kale, and Aditi Raghunathan. Mitigating modal imbalance in multimodal reasoning, 2025. URL https://arxiv.org/abs/2510.02608.

[46] Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection, 2025. URL https://arxiv.org/abs/2501.00654.

[47] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning, 2024. URL https://arxiv.org/abs/2402.04333.

[48] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. URL https://arxiv.org/abs/2402.11690.

[49] Yichen Yan, Zhaoyi Yuan, Jiajun Pan, Xiu Tang, Gongsheng Yuan, Xiaoling Gu, Jinpeng Chen, Sai Wu, Ke Chen, Mingli Song, et al. Data quality management for large vision-language models: Issues, techniques, and prospects. *Authorea Preprints*, 2025.

[50] Yichen Yan, Ming Zhong, Qi Zhu, Xiaoling Gu, Jinpeng Chen, and Huan Li. Coido: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization, 2025. URL https://arxiv.org/abs/2510.17847.

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

[52] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

[53] Zhihan Zhou, Feng Hong, Jiaan Luo, Jiangchao Yao, Dongsheng Li, Bo Han, Ya Zhang, and Yanfeng Wang. Learning to instruct for visual instruction tuning, 2025. URL https://arxiv.org/abs/2503.22215.

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

Appendix

A Related work

Existing VLLM data selection methods can be categorized into score-based and clustering-based methods. In this section, we present a detailed overview of these methods.

A.1 Score-based methods

Score-based methods assign an importance score to each sample and perform selection accordingly. Early methods rely on simple proxies such as prediction error or perplexity (e.g., EL2N [38], MP [31]). More recent methods incorporate gradient-based signals to estimate sample contribution or influence, including Self-Filter [8], LESS [47], ICONS [46], and TIVE [26]. Other variants further consider diversity or grouped sampling (e.g., COIDO [50], MLLM-Selector [29]). Most score-based methods require training proxy models, incurring substantial overhead for large-scale multimodal data. Training-free alternatives such as CLIP-Score [13] avoid this cost but are limited to coarse-grained semantic alignment [37]. In contrast, our method leverages a frozen VLLM to model conditional shifts via relative judgment changes, enabling fine-grained semantic relationships modeling without proxy training.

A.2 Clustering-based methods

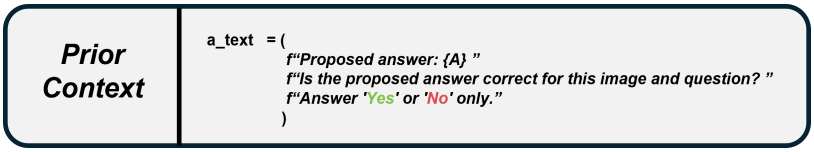
Clustering-based methods promote diversity by modeling relationships among samples, typically by performing clustering in the embedding space of a specific model layer. Early works such as Self-Sup [41] and SemDeDup [1] remove redundant samples based on cluster structure or intra-cluster similarity, while D2 Pruning [30] models soft neighborhoods via graph message passing. More recent methods exploit richer or cross-modal representations. COINCIDE [18] clusters joint representations from multiple layers to capture concept–skill structure, and XMAS [36] clusters samples based on cross-modal attention trajectories. ARDS [51] further improves robustness through worst-case subgroup construction and weighted sampling. Existing clustering-based methods typically involve additional operations after feature extraction, and some still suffer from the overhead of training proxy models. In contrast, our method is more cost-efficient, as it directly leverages a frozen VLLM without requiring any additional post-clustering procedures or proxy training.

B Distributional perspectives on data selection

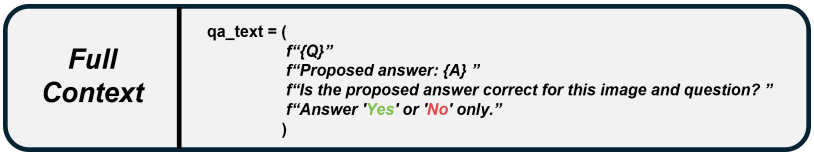
We emphasize that this perspective is introduced as an interpretative lens, rather than a formal objective or a theoretical derivation of our method. In practice, training data are rarely sampled directly from the real-world distribution $\mathcal{P}_{\text{real}}$, but instead from a constructed distribution $\mathcal{P}_{\text{train}}$ formed through data integration or filtering. Building on recent advances in data selection [10], the training of VLLMs can be viewed as an interaction among three components:

- **Generator ($\mathcal{G}_\theta$):** The VLLM learner, whose parameters θ are optimized via gradient descent to model conditional distributions.
- **Evaluator ($\mathcal{D}_\phi$):** A parameterized model or scoring function that selects or reweights training samples from raw data sources.
- **Target Distribution ($\mathcal{P}_{real}$):** The real-world distribution used to assess generalization performance.

Under this formulation, the generalization error L_{test} depends not only on the training set size N, but also on the interaction between the Generator's modeling capacity and the Evaluator's ability to align $\mathcal{P}_{\text{train}}$ with $\mathcal{P}_{\text{real}}$. Prior theory suggests a phase transition: when the Generator is sufficiently expressive and the Evaluator effectively filters for informative, distribution-consistent samples, strong



(a) Prior context prompt.



(b) Full context prompt.

Figure 5: Prompt templates for CVS computation.

generalization can be achieved with substantially reduced data, corresponding to a "less is more" regime.

C CVS prompt templates

As illustrated in Figure 5, we provide the complete prompt templates used to obtain binary answer validity judgments from a frozen VLLM. The model is instructed to output only "Yes" or "No".

D Complementarity of affirmation and rejection shifts

In the main paper, we use Conditional Affirmation Shift (CVS_{YES}) and Conditional Rejection Shift (CVS_{NO}) as two separate constraints. Since the evaluator predicts over $\{\text{Yes}, \text{No}\}$, one may wonder whether these two quantities are redundant, and whether they can be replaced by a single coupled score. We therefore conduct an ablation on Vision-Flan under the 10% budget setting to examine the necessity of keeping them separate.

Specifically, we compare three types of variants: using only CVS_{YES}, using only CVS_{NO}, and replacing the two constraints with a coupled log-odds change:

$$\Delta = \log \frac{P(\text{Yes} \mid I, Q, A)}{P(\text{No} \mid I, Q, A)} - \log \frac{P(\text{Yes} \mid I, A)}{P(\text{No} \mid I, A)}. \tag{11}$$

For the coupled score, we evaluate three selection strategies, denoted as $\Delta_{P_{\text{Min}}}$, $\Delta_{P_{\text{Random}}}$, and $\Delta_{P_{\text{Max}}}$.

As shown in Table 4, both individual constraints provide useful signals, with CVS_{YES} and CVS_{NO} achieving 100.59 and 101.89 ARP, respectively. However, combining the two shifts achieves the best performance, improving the ARP to 103.45. In contrast, the coupled log-odds variants are consistently inferior to the separated formulation, with the best coupled variant reaching only 102.11 ARP. These results suggest that although the Yes/No probabilities are naturally correlated, their shifts are not redundant in practice. Separating CVS_{YES} and CVS_{NO} better preserves the distinction between answer support and semantic conflict, which leads to more effective filtering than collapsing them into a single scalar.

E Probability normalization and shift sign constraints

CVS uses the original probabilities of the YES and NO tokens over the full vocabulary. These probabilities are not renormalized over the two verdict labels and need not sum to one. Introducing the question can increase or decrease both probabilities by shifting probability mass to or from other tokens. A positive CVS_{YES} does not necessarily imply a negative CVS_{NO}. This implication would hold if the probabilities were normalized over the two labels.

Table 4: Effect of separating affirmation and rejection constraints.

Score	ARP
$\mathrm{CVS}_{\mathrm{YES}}$	100.59
$\mathrm{CVS}_{\mathrm{NO}}$	101.89
$\Delta_{P_{\mathrm{Min}}}$	94.53
$\Delta_{P_{\mathrm{Random}}}$	101.62
$\Delta_{P_{\mathrm{Max}}}$	102.11
$\mathrm{CVS}_{\mathrm{YES}} + \mathrm{CVS}_{\mathrm{NO}}$	**103.45**

Table 5: Prompt sensitivity analysis of CVS.

Method	ARP
Random	95.30
{YES, NO}	103.45
{CORRECT, INCORRECT}	103.69
{RIGHT, WRONG}	103.00
Can...question?	**104.54**

The joint condition $\mathrm{CVS}_{\mathrm{YES}} > 0$ and $\mathrm{CVS}_{\mathrm{NO}} < 0$ requires the question to increase answer support and reduce rejection according to the evaluator. Requiring both conditions imposes a stricter consistency criterion than either condition alone. The opposite combination, $\mathrm{CVS}_{\mathrm{YES}} < 0$ and $\mathrm{CVS}_{\mathrm{NO}} > 0$, indicates that the question weakens answer support and strengthens rejection. The validity filter excludes these samples before difficulty ranking.

F Prompt sensitivity analysis

CVS relies on evaluator probabilities assigned to binary answer tokens. Therefore, the resulting scores may be affected by the choice of answer verbalizers or by the specific wording of the evaluator prompt. To examine the robustness of CVS to these design choices, we conduct a prompt sensitivity analysis on Vision-Flan under the 10% data budget.

We consider two types of variants. First, we keep the prompt template unchanged and vary only the binary answer tokens, replacing {YES, NO} with {CORRECT, INCORRECT} and {RIGHT, WRONG}. Second, we keep the answer tokens as {YES, NO} but modify the prompt wording by replacing the original template with the question "Can the proposed answer be regarded as correct for this image and question?" All other components of the selection pipeline are kept unchanged.

As shown in Table 5, all CVS variants substantially outperform random selection. More importantly, changing the answer verbalizers leads to only minor performance variation: {YES, NO}, {CORRECT, INCORRECT}, and {RIGHT, WRONG} achieve 103.45, 103.69, and 103.00 ARP, respectively. The alternative prompt template further achieves 104.54 ARP. These results indicate that CVS is not overly sensitive to a particular choice of binary answer tokens or prompt wording. Instead, its effectiveness mainly comes from the conditional probability shift induced by the image–question–answer context, rather than from a specific prompt instantiation.

G Performance at higher sampling ratios

We evaluate CVS on Vision-Flan using 20%, 25%, and 30% of the data, with all other settings matching the main experiments. Table 6 reports ARP for CVS and random sampling. CVS outperforms random sampling at all three ratios, showing that its advantage extends to larger data budgets.

Table 6: ARP on Vision-Flan at higher sampling ratios. The best result in each column is highlighted in bold.

Method	20%	25%	30%
Random	96.40	96.95	97.35
CVS	**98.82**	**101.57**	**100.33**

Table 7: Hyperparameter configurations.

Method	Settings
CLIP-Score	high score selected
EL2N	medium score selected
SemDeDup	$K : 5,000$
D2 Pruning	$k : 5, \gamma_r : 0.4, \gamma_f : 1.0$
COINCIDE	$K : 5,000, \tau : 0.1$
XMAS	$K : 1,000, T : 7$

H Dataset details

H.1 Vision-Flan

Vision-Flan [48] is a high-quality human-annotated visual instruction tuning dataset designed to systematically enhance the generalization capability of vision-language models across diverse visual tasks. The dataset is constructed entirely from publicly available academic vision datasets and covers 187 finely grained visual tasks, each accompanied by expert-written and rigorously validated task instructions. This design significantly reduces synthetic bias and the risk of hallucination.

Prior studies have demonstrated that Vision-Flan offers substantial benefits in improving overall visual understanding while mitigating catastrophic forgetting, making it one of the publicly available VIT datasets with the widest task diversity and highest annotation reliability to date. Each task contains approximately 1,000 samples, resulting in a total of 186K samples, all of which are used in our experiments.

H.2 The Cauldron

The Cauldron [17] is a large-scale, high-quality visual instruction tuning dataset specifically constructed for instruction tuning of general-purpose vision-language large models. It integrates 50 vision-language datasets spanning a wide range of task types, including visual question answering, image captioning, text transcription, document and table understanding, chart reasoning, visual logical reasoning, and code generation across diverse and complex scenarios.

All constituent datasets are unified into a standardized question–answer format, and rigorous deduplication and evaluation-set decontamination procedures are applied during dataset construction to effectively mitigate the risk of data leakage. Given the substantial scale of the Cauldron dataset, to facilitate broad and fair evaluation of multiple baseline methods under a reasonable computational budget, we construct a representative subset of 100K samples by random sampling, while strictly preserving the original task distribution.

I Hyperparameter settings for baseline methods

For all baseline methods other than XMAS, we use the hyperparameter settings reported in the COINCIDE paper. For XMAS, we follow the hyperparameter configuration described in its paper, as summarized in Table 7.

J Rationale for the random sampling baseline

Random sampling is also a strong baseline and provides a consistent reference without additional selection criteria. Under matched data budgets and training conditions, it helps assess the effects of CVS selection strategies and model choices. Switching to the strongest baseline at each sampling ratio would make these effects harder to interpret. We therefore primarily use random sampling in analysis and ablation studies, while comparing against competing methods in the main experiments.

K Benchmark details

GQA GQA [14] is a visual question answering benchmark designed for real-world visual reasoning. It automatically generates questions with explicit compositional structure from scene graphs and provides explicit annotations of multi-step reasoning processes. By systematically balancing the distributions of questions and answers, GQA substantially reduces language bias, enabling more reliable evaluation of models' compositional vision-language reasoning capabilities.

VizWiz VizWiz [12] is a visual question answering benchmark targeting real-world assistive scenarios, with data collected from images captured by blind users using mobile phones along with their spoken questions. The benchmark contains a large number of low-quality images and unanswerable questions, and explicitly incorporates answerability prediction into its evaluation protocol, enabling assessment of VQA models' robustness in real-world, unconstrained environments.

TextVQA TextVQA [40] is a visual question answering benchmark focused on understanding text in images, comprising approximately 28K natural scene images and 45K questions that require reading and reasoning over textual content to answer. The benchmark emphasizes open-vocabulary answers and OCR-based copying mechanisms, and is designed to evaluate models' text reading and multimodal reasoning abilities in real-world scenarios.

ScienceQA-IMG ScienceQA-IMG [27] is a subset of ScienceQA that contains only questions with image-based context, including both natural scene images and instructional diagrams. It requires models to perform multiple-choice scientific question answering by integrating visual information. This subset emphasizes image understanding and multi-step scientific reasoning, and provides ground-truth answers along with corresponding explanation annotations, making it suitable for evaluating multimodal reasoning performance.

MME MME [21] is a comprehensive benchmark designed to systematically assess the perception and reasoning capabilities of multimodal large language models. It covers a diverse set of evaluation dimensions, including object recognition, attribute understanding, spatial reasoning, OCR, commonsense reasoning, and visual reasoning, spanning both perceptual-level and cognitive-level skills. MME adopts a unified evaluation protocol with carefully curated prompts and scoring criteria, enabling fine-grained and reliable comparison of multimodal models across heterogeneous abilities.

MMBench MMBench [24] is a multi-capability evaluation benchmark for large-scale vision-language models, consisting of approximately 3K multiple-choice questions that cover 20 fine-grained ability dimensions, spanning both perceptual and reasoning skills. The benchmark employs strict data quality control and a CircularEval evaluation strategy, enabling more robust and reliable comparison of models' comprehensive multimodal capabilities.

AI2D AI2D [15] is a benchmark for scientific diagram understanding, consisting of over 5K diagrams drawn from elementary-level science textbooks and approximately 15K multiple-choice questions. The benchmark emphasizes modeling entities, relations, and structural layouts within diagrams, and is designed to evaluate models' diagram parsing and relation-based reasoning abilities.

ChartQA ChartQA [32] is a question answering benchmark for real-world chart understanding, featuring charts derived from real data sources and human-authored questions that emphasize joint visual and logical reasoning. The benchmark supports open-vocabulary answers and requires models to simultaneously comprehend visual chart elements and underlying numerical relationships, making it suitable for evaluating complex chart reasoning capabilities.

DocVQA DocVQA [33] is a visual question answering benchmark for document image understanding, comprising approximately 12K real-world document images and 50K text-based questions. The benchmark emphasizes joint modeling of document layout, structure, and dense textual content, and

is designed to evaluate models' document-level reading comprehension and information localization abilities.

InfoVQA InfoVQA [34] is a visual question answering benchmark for infographic understanding, comprising approximately 5.4K infographic images and 30K questions that require joint modeling of text, layout, graphics, and data visualization elements. The benchmark emphasizes reasoning skills such as counting, ordering, and basic arithmetic, and is designed to evaluate models' performance on complex document-level vision-language understanding tasks.

MMStar MMStar [7] is a multimodal evaluation benchmark that emphasizes visual indispensability, consisting of 1,500 carefully human-filtered questions for which correct answers must rely on visual information. The benchmark spans 6 core capabilities and 18 fine-grained dimensions, and is designed to more faithfully evaluate large models' multimodal understanding and reasoning abilities, while explicitly reducing the impact of textual shortcuts and data leakage.

OCRBench OCRBench [25] is a comprehensive benchmark for evaluating the OCR capabilities of large-scale multimodal models, covering five task categories: text recognition, scene text VQA, document VQA, key information extraction, and handwritten formula recognition. The benchmark consists of 1,000 human-curated and manually verified question–answer pairs, and is designed to systematically analyze models' actual strengths and limitations across diverse text-centric visual tasks.

L Benchmark result details

This section summarizes the detailed results of all experiments in this paper. Table 9 reports the complete results on Vision-Flan for different data selection methods using 5%, 10%, and 15% data subsets. Table 10 presents the complete results on The Cauldron for different data selection methods using 10%, 20%, 30%, and 50% data subsets. Table 11 shows the impact of different CVS score selection strategies (Low / High / No) on model performance on Vision-Flan. Tables 12 and 13 compare the performance of CVS under different evaluator architectures and different evaluator scale settings, respectively. Table 14 and 15 report robustness evaluation results when changing the target model while keeping the evaluator fixed. Finally, Table 16 presents an ablation study that removes the visual anchoring term, aiming to analyze the role of visual information in the design of CVS.

M Examples across CVS score ranges

In this section, we provide representative examples across different CVS score ranges, including Low, High, and No samples. The corresponding examples are shown in Figures 6, 7, and 8, respectively, to qualitatively illustrate the characteristics of samples within each score range.

N Validity filtering and difficulty ranking

CVS follows a two-stage procedure. First, validity filtering retains samples satisfying $\mathrm{CVS}_{\mathrm{YES}} > 0$ and $\mathrm{CVS}_{\mathrm{NO}} < 0$. These conditions indicate that introducing the question consistently supports answer validity according to the evaluator by increasing answer support and reducing rejection.

Second, difficulty ranking prioritizes samples with the lowest positive $\mathrm{CVS}_{\mathrm{YES}}$ within the filtered pool under the data budget. A smaller positive shift indicates that the question supports answer validity, but only by a small relative margin. CVS treats these samples as hard positives and prioritizes them to encourage stronger integration of visual content and linguistic instructions. The shift signs determine admissibility, while the affirmation-shift magnitude ranks difficulty among the retained samples.

O Manual analysis of samples across selection stages

We randomly sample 200 examples from each of three groups on Vision-Flan: samples excluded by validity filtering, samples not selected by subsequent difficulty ranking, and the final selected subset. The second group contains candidates that pass the validity filter but are not prioritized under the data budget because of their higher $\mathrm{CVS}_{\mathrm{YES}}$ scores. Two PhD students and one master's student in computer science independently classify the examples as language shortcuts, labeling errors, or correct samples without language shortcuts. Disagreements are resolved by majority vote.

Table 8 shows that the first stage primarily removes samples with labeling errors (58%). The second stage excludes language shortcuts (20%) and correct samples (67%), many of which require only simple visual judgments. In the final subset, 98% of the inspected samples are correct, with many requiring careful visual reasoning. No language shortcuts are identified in this inspected group.

The first stage also excludes 40% correct samples, indicating that validity filtering is imperfect. These exclusions may reflect limitations of the frozen evaluator. CVS aims to retain useful samples within a fixed data budget rather than recover every valid example.

Table 8: Manual analysis of samples across CVS selection stages.

Stage	Language shortcuts (%)	Label errors (%)	Correct (%)
1	2	58	40
2	20	13	67
Final	0	2	98

P Discussion on absolute verdict confidence

CVS primarily relies on relative verdict shifts rather than directly ranking samples by the absolute probability $P(\mathrm{Yes} \mid I, Q, A)$. This design does not imply that the absolute verdict confidence is unimportant. In principle, $P(\mathrm{Yes} \mid I, Q, A)$ reflects the evaluator's confidence in the correctness of the answer under the full multimodal context, and extremely low confidence may indicate ambiguous, noisy, or low-quality supervision. However, we find that absolute confidence values are less reliable as a universal selection criterion across heterogeneous visual instruction tasks, since different tasks, answer formats, and prompt styles can induce substantially different confidence ranges in the evaluator. As a result, a fixed absolute-confidence threshold may either over-filter valid but challenging samples or retain easy samples with high linguistic priors.

In contrast, CVS measures the conditional change in the evaluator's judgment caused by introducing the question Q under the same visual context. This relative formulation provides a more task-stable signal of whether the question contributes useful information for validating the answer. Moreover, low-quality or mathematically ambiguous samples are unlikely to be favored by the CVS protocol. If a sample is noisy, unclear, or semantically inconsistent, introducing the question typically leads to unstable verdict changes and often fails to satisfy the joint filtering condition $CVS_{\mathrm{Yes}} > 0$ and $CVS_{\mathrm{No}} < 0$. Conversely, if a sample is overly easy, its CVS_{Yes} tends to be large, indicating that the evaluator can validate the answer with high confidence once the question is provided; such samples are not prioritized by our low-score strategy. Therefore, CVS favors samples of moderate difficulty: they exhibit a small positive affirmation shift together with a negative rejection shift, suggesting that the question provides meaningful but non-trivial guidance for vision-language reasoning. This makes them more suitable for effective visual instruction tuning.

Q Connection to conditional information gain

CVS is information-gain inspired, but it is not intended to be an exact estimator of the conditional mutual information $I(Q; Y \mid I)$. Estimating $I(Q; Y \mid I)$ would require access to the true data

distribution and marginalization over possible questions and verdicts, which is generally infeasible for large-scale visual instruction data. Instead, CVS uses the conditional likelihood ratio

$$\log \frac{P(Y \mid I, Q, A)}{P(Y \mid I, A)} \tag{12}$$

as a computable sample-level proxy for measuring how much the question Q changes the evaluator's answer-validity judgment under the same visual context I.

Therefore, the goal of CVS is not to recover the absolute value of conditional mutual information, but to obtain a stable ranking signal for data selection. A positive affirmation shift indicates that the question provides additional support for the answer, while a positive rejection shift indicates possible semantic conflict. By combining these two relative shifts, CVS operationally distinguishes informative cross-modal samples from easy, noisy, or misaligned samples.

Table 9: Main experimental results on Vision-Flan. We compare different data selection methods on Vision-Flan using 5%, 10%, and 15% data subsets. The best result among subset-selection methods in each column is highlighted in bold.

(a) Detailed results of different methods on Vision-Flan using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	42.93	55.79	36.54	60.34	1098.02	36.40	26.19	37.82	15.96	16.46	19.99	34.77	27.80	92.96
XMAS	40.38	54.15	37.06	62.42	705.48	39.95	39.95	41.22	15.40	17.98	19.80	34.37	27.50	93.46
COINCIDE	42.15	53.48	**38.00**	55.03	1000.71	29.12	41.24	36.40	16.64	17.89	17.79	33.51	28.70	92.26
D2Prune	41.02	55.99	37.47	60.59	928.95	36.94	44.93	38.54	16.60	17.68	20.58	33.47	27.90	95.61
EL2N	**43.26**	54.93	35.94	54.34	898.77	39.78	18.90	34.03	15.80	16.23	18.36	34.13	28.20	88.74
SemDeDup	43.09	55.72	37.31	63.11	**1115.77**	41.32	48.71	41.16	**17.80**	19.02	19.42	33.09	27.80	**99.45**
CLIP Score	39.24	54.77	31.80	57.71	549.88	38.06	40.38	37.63	15.92	16.78	19.69	33.75	**29.20**	89.92
SCALSELECT	39.52	53.96	36.08	62.87	838.18	38.49	43.64	**43.13**	15.28	17.67	19.07	34.01	27.50	93.86
PRISM	29.97	52.06	26.73	49.33	1106.50	35.13	16.67	29.99	17.00	**19.56**	19.35	35.26	27.90	86.08
CVS (Ours)	38.42	**57.46**	37.98	**63.41**	800.81	**45.88**	**52.66**	39.44	15.12	18.48	**20.74**	**35.59**	27.80	97.82

(b) Detailed results of different methods on Vision-Flan using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	43.31	56.02	36.13	60.92	1152.77	39.61	31.88	44.50	**16.52**	17.27	18.08	34.36	27.90	95.30
XMAS	42.73	55.82	36.55	64.40	1008.41	45.19	54.38	43.95	16.20	18.39	18.19	34.28	28.50	99.31
COINCIDE	43.33	57.38	38.70	60.83	1032.00	43.56	41.24	40.16	15.84	18.96	19.25	34.27	29.00	97.93
D2Prune	43.47	56.40	37.81	63.01	1146.50	**47.68**	51.80	43.01	16.44	17.38	18.78	34.73	29.00	100.49
EL2N	**44.08**	56.99	38.10	47.99	1088.62	39.52	31.70	29.92	15.64	18.78	18.78	34.94	28.30	92.67
SemDeDup	43.93	56.39	37.20	58.65	**1236.25**	47.51	45.96	39.31	15.72	18.89	17.58	36.00	28.50	98.97
CLIP Score	41.02	53.63	38.11	59.84	725.09	34.28	30.50	35.56	16.36	20.29	22.72	34.90	**29.10**	94.13
SCALSELECT	42.73	56.26	37.68	60.88	1101.46	44.42	**54.81**	44.49	14.84	18.29	16.71	33.56	27.80	97.94
PRISM	35.21	48.80	37.87	61.97	751.61	38.49	44.16	**44.88**	16.44	**23.76**	**23.37**	35.92	28.80	98.99
CVS (Ours)	39.49	**57.66**	**38.99**	**66.78**	930.81	46.13	54.30	43.98	16.24	21.18	21.72	**37.70**	28.70	**103.45**

(c) Detailed results of different methods on Vision-Flan using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	43.19	56.33	35.41	59.59	1230.04	42.44	38.06	41.06	15.52	16.26	16.26	32.27	27.60	93.82
XMAS	44.68	56.52	37.19	61.18	1012.67	40.72	46.99	44.62	15.84	17.83	17.88	34.95	27.50	97.13
COINCIDE	44.28	56.44	37.52	55.73	1221.20	45.96	39.26	37.50	17.24	19.03	18.32	35.59	28.80	98.39
D2Prune	43.96	57.16	38.40	61.63	1059.64	44.67	49.48	44.66	15.80	17.62	18.60	34.13	28.30	98.94
EL2N	45.81	57.45	38.52	53.69	1105.20	46.22	39.18	43.39	14.68	19.06	17.35	34.56	28.40	96.87
SemDeDup	**50.00**	51.57	37.30	59.69	**1247.46**	43.90	45.45	43.13	17.04	18.66	18.93	35.24	28.00	100.35
CLIP Score	41.33	55.10	38.44	**65.25**	640.09	39.60	41.67	**48.41**	16.68	20.19	21.96	37.57	**29.30**	99.24
SCALSELECT	43.15	57.90	38.47	57.56	1062.93	46.91	43.99	36.40	15.96	18.82	16.69	35.53	27.70	96.65
PRISM	39.25	52.24	38.62	61.23	770.61	41.67	41.07	43.78	**17.96**	21.12	21.00	36.22	29.10	98.64
CVS (Ours)	43.46	**57.93**	**40.57**	63.46	1123.40	**47.17**	**51.03**	42.23	16.00	**21.36**	**21.98**	**37.64**	29.00	**104.85**

Table 10: Main experimental results on The Cauldron. We compare different data selection methods on The Cauldron using 10%, 20%, 30%, and 50% data subsets. The best result in each column is highlighted in bold.

(a) Detailed results of different methods on The Cauldron using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	46.76	58.16	49.13	78.33	1333.04	57.73	65.72	64.57	12.80	27.52	15.88	40.38	35.10	100.00
Random	44.10	**48.73**	42.82	66.23	1038.67	45.73	57.19	50.91	10.24	20.55	14.03	34.08	30.10	83.54
XMAS	**44.56**	41.38	44.10	**69.11**	1104.39	48.37	56.27	52.59	10.32	20.62	14.14	34.26	30.40	**84.17**
COINCIDE	43.85	45.39	43.40	67.17	**1118.15**	48.48	54.60	50.39	10.40	20.31	14.39	34.15	30.20	83.94
D2Prune	43.06	43.29	43.35	68.91	1009.67	47.62	56.76	52.59	10.12	19.68	13.89	34.60	28.80	82.66
EL2N	43.55	41.20	43.94	68.95	1072.84	48.39	55.03	50.26	**10.68**	20.78	**14.80**	**35.10**	30.50	84.10
SemDeDup	43.02	45.04	43.24	69.00	1003.83	**48.65**	**57.96**	51.33	9.40	20.43	14.48	33.87	30.10	83.18
CLIP Score	44.08	38.73	**44.13**	67.42	1072.95	48.05	55.29	**56.56**	9.08	19.90	13.67	34.82	30.40	82.64
CVS (Ours)	43.58	40.58	43.90	67.64	1000.09	48.25	56.88	53.79	9.64	**21.15**	14.45	34.77	**31.40**	83.53

(b) Detailed results of different methods on The Cauldron using a 20% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	46.76	58.16	49.13	78.33	1333.04	57.73	65.72	64.57	12.80	27.52	15.88	40.38	35.10	100.00
Random	44.50	49.20	44.78	68.27	1116.78	50.80	58.47	56.09	10.64	22.23	15.23	35.23	**31.90**	**87.96**
XMAS	44.20	42.83	**45.80**	71.24	1086.13	49.83	57.30	55.18	10.60	22.77	**15.58**	35.14	30.60	86.97
COINCIDE	**45.04**	48.11	45.05	70.96	**1152.90**	50.03	56.92	53.53	9.96	22.04	14.96	33.45	30.60	86.63
D2Prune	43.83	43.58	44.33	**71.63**	1131.72	51.32	57.70	55.21	**11.48**	21.83	14.44	31.78	30.40	86.35
EL2N	43.67	43.36	44.45	70.49	1101.89	50.55	57.10	55.86	10.76	22.27	14.62	33.73	30.90	86.20
SemDeDup	44.11	**49.40**	44.89	70.49	1041.51	49.68	56.67	55.11	10.36	22.44	14.87	34.44	31.50	86.73
CLIP Score	44.74	37.79	45.11	71.28	1144.76	**51.74**	**59.50**	**59.71**	10.00	21.90	14.89	**37.07**	30.70	87.12
CVS (Ours)	44.57	45.84	45.03	70.51	1035.29	50.23	56.19	57.09	10.56	**22.92**	15.52	34.64	31.00	87.09

(c) Detailed results of different methods on The Cauldron using a 30% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	46.76	58.16	49.13	78.33	1333.04	57.73	65.72	64.57	12.80	27.52	15.88	40.38	35.10	100.00
Random	45.15	44.57	45.66	72.23	1188.38	51.49	58.22	55.61	**11.08**	23.07	15.62	34.45	32.00	88.97
XMAS	44.89	47.14	**46.60**	72.78	1117.63	**52.92**	**61.51**	58.52	10.92	24.02	15.06	35.76	32.50	90.24
COINCIDE	45.11	50.12	45.29	**73.61**	**1235.28**	52.69	60.36	57.93	10.84	**25.39**	15.55	34.35	31.80	**91.14**
D2Prune	44.18	44.52	45.83	73.07	1126.04	51.75	60.28	57.78	10.00	22.45	14.86	**35.90**	32.20	88.22
EL2N	**45.83**	**50.83**	45.16	71.73	1213.79	51.91	60.45	57.64	10.72	23.35	14.77	33.73	30.90	89.56
SemDeDup	44.34	42.59	44.98	72.08	1164.76	52.34	60.62	56.93	**11.08**	23.33	15.04	35.80	31.70	88.85
CLIP Score	45.31	39.77	45.85	71.38	1235.07	52.77	61.39	**60.52**	10.56	22.86	15.09	35.68	32.30	89.38
CVS (Ours)	44.92	50.73	45.14	71.36	1095.65	49.89	59.97	59.11	10.64	23.48	**15.78**	35.41	**32.70**	89.72

(d) Detailed results of different methods on The Cauldron using a 50% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	46.76	58.16	49.13	78.33	1333.04	57.73	65.72	64.57	12.80	27.52	15.88	40.38	35.10	100.00
Random	45.38	49.75	46.60	73.97	1136.19	53.47	61.40	60.50	11.32	24.69	15.06	36.25	32.70	91.75
XMAS	45.44	44.92	**46.72**	75.11	1223.29	54.12	61.94	59.23	11.56	**25.24**	15.33	35.20	33.10	92.07
COINCIDE	**46.55**	51.45	46.60	75.25	1236.37	**55.09**	62.51	60.33	11.32	24.14	14.92	35.91	33.60	93.11
D2Prune	45.16	**54.69**	45.96	73.57	1258.27	53.63	62.16	59.26	**12.04**	24.70	15.10	35.99	33.70	**93.52**
EL2N	45.21	47.63	46.30	74.41	1222.04	53.46	61.65	59.16	10.52	24.69	15.19	35.10	33.70	91.39
SemDeDup	45.39	47.50	46.68	**75.26**	**1281.12**	54.92	62.25	59.36	11.32	24.67	**15.47**	36.47	32.50	92.78
CLIP Score	45.98	45.73	46.46	73.07	1215.97	53.20	**63.11**	60.81	11.96	24.27	14.83	37.22	32.10	92.07
CVS (Ours)	44.99	45.33	45.86	75.22	1181.08	54.10	62.89	**61.69**	10.72	24.88	14.95	**38.25**	**34.40**	92.16

Table 11: Detailed results on Vision-Flan for the effect of CVS score ranges. We evaluate different CVS score selection strategies (Low / High / No) on the Vision-Flan using data subsets of 5%, 10%, and 15%. The best result in each column is highlighted in bold.

(a) Detailed results of different CVS score selection strategies on Vision-Flan using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**42.93**	55.79	36.54	60.34	**1098.02**	36.40	26.19	37.82	**15.96**	16.46	19.99	34.77	27.80	92.96
No	24.31	53.87	29.59	53.20	564.60	31.70	38.66	27.14	14.80	13.43	15.22	35.02	**27.90**	79.16
High	35.07	51.09	30.75	62.27	488.55	39.69	44.16	**45.95**	15.12	14.55	19.77	31.38	23.80	87.16
Low	38.42	**57.46**	**37.98**	**63.41**	800.81	**45.88**	**52.66**	39.44	15.12	**18.48**	**20.74**	**35.59**	27.80	**97.82**

(b) Detailed results of different CVS score selection strategies on Vision-Flan using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**43.31**	56.02	36.13	60.92	**1152.77**	39.61	31.88	44.50	**16.52**	17.27	18.08	34.36	27.90	95.30
No	34.72	54.65	36.96	48.88	870.97	32.90	40.89	22.63	15.52	17.80	20.56	37.18	28.60	89.48
High	38.09	55.15	33.52	59.79	533.81	45.79	36.00	**44.66**	15.44	15.81	18.22	32.50	25.20	89.05
Low	39.49	**57.66**	**38.99**	**66.78**	930.81	**46.13**	**54.30**	43.98	16.24	**21.18**	**21.72**	**37.70**	**28.70**	**103.45**

(c) Detailed results of different CVS score selection strategies on Vision-Flan using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	43.19	56.33	35.41	59.59	**1230.04**	42.44	38.06	41.06	15.52	16.26	16.26	32.27	27.60	93.82
No	36.61	55.70	37.60	62.42	668.22	40.21	39.69	39.25	**16.88**	16.41	18.21	36.19	**29.30**	92.76
High	40.32	54.78	33.47	63.26	830.61	45.70	**53.01**	**53.69**	15.44	16.26	17.12	35.81	26.40	96.00
Low	**43.46**	**57.93**	**40.57**	**63.46**	1123.40	**47.17**	51.03	42.23	16.00	**21.36**	**21.98**	**37.64**	29.00	**104.85**

Table 12: Detailed results on Vision-Flan under different evaluator architectures. We compare different evaluator architectures on Vision-Flan using data subsets of 5%, 10%, and 15%. Qwen, InternVL, and LLaVA-OV correspond to Qwen2.5-VL-7B-Instruct, InternVL3-8B, and LLaVA-OneVision-1.5-4B-Instruct, respectively. The best result among subset-selection methods in each column is highlighted in bold.

(a) Detailed results of different evaluator architectures on Vision-Flan using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**42.93**	55.79	36.54	60.34	**1098.02**	36.40	26.19	37.82	**15.96**	16.46	19.99	34.77	27.80	92.96
CVS(Qwen)	38.42	57.46	**37.98**	63.41	800.81	45.88	**52.66**	39.44	15.12	**18.48**	**20.74**	35.59	27.80	**97.82**
CVS(InternVL)	36.93	**57.97**	34.42	60.09	839.81	**46.48**	43.73	39.67	14.56	16.34	18.24	**38.27**	**28.70**	93.89
CVS(LLaVA-OV)	39.01	54.20	34.99	**64.51**	1084.98	35.05	45.27	**47.96**	14.76	17.50	20.58	31.17	27.20	95.39

(b) Detailed results of different evaluator architectures on Vision-Flan using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**43.31**	56.02	36.13	60.92	**1152.77**	39.61	31.88	**44.50**	16.52	17.27	18.08	34.36	27.90	95.30
CVS(Qwen)	39.49	**57.66**	38.99	**66.78**	930.81	46.13	**54.30**	43.98	16.24	21.18	**21.72**	37.70	28.70	**103.45**
CVS(InternVL)	36.71	54.36	**39.91**	61.82	631.84	**48.72**	43.38	40.16	**16.56**	18.75	19.06	**39.08**	29.00	96.68
CVS(LLaVA-OV)	37.02	54.21	39.65	60.34	871.82	40.71	34.02	39.12	15.36	**21.52**	21.11	35.31	**30.30**	96.59

(c) Detailed results of different evaluator architectures on Vision-Flan using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	43.19	56.33	35.41	59.59	**1230.04**	42.44	38.06	41.06	15.52	16.26	16.26	32.27	27.60	93.82
CVS(Qwen)	**43.46**	**57.93**	40.57	**63.46**	1123.40	47.17	**51.03**	**42.23**	16.00	21.36	**21.98**	**37.64**	29.00	**104.85**
CVS(InternVL)	35.20	57.04	39.61	61.58	550.58	**49.91**	50.00	41.48	**17.00**	20.44	19.65	37.32	29.10	98.49
CVS(LLaVA-OV)	39.05	57.20	**40.87**	60.78	524.28	45.36	44.50	41.22	15.60	**21.73**	21.52	36.30	**30.00**	98.56

Table 13: Detailed results under different evaluator scales on Vision-Flan. We compare evaluators of different scales within the Qwen2.5-VL family, where CVS (7B) and CVS (3B) denote using Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct as evaluators, respectively. Experiments are conducted using 5%, 10%, and 15% data subsets. The best result in each column is highlighted in bold.

(a) Detailed results under different evaluator scales on Vision-Flan using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
CVS (7B)	38.42	**57.46**	**37.98**	**63.41**	**800.81**	45.88	**52.66**	**39.44**	15.12	**18.48**	**20.74**	35.59	27.80	**97.82**
CVS (3B)	**39.17**	53.22	34.89	62.33	500.00	**46.22**	43.56	39.18	**15.20**	15.48	17.74	**37.50**	**28.60**	91.24

(b) Detailed results under different evaluator scales on Vision-Flan using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
CVS (7B)	39.49	**57.66**	38.99	**66.78**	**930.81**	**46.13**	**54.30**	**43.98**	16.24	**21.18**	**21.72**	37.70	**28.70**	**103.45**
CVS (3B)	**40.13**	53.27	**39.65**	63.16	513.92	43.47	50.43	41.39	**16.84**	18.15	18.28	**39.60**	28.40	96.24

(c) Detailed results under different evaluator scales on Vision-Flan using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
CVS (7B)	**43.46**	**57.93**	**40.57**	**63.46**	**1123.40**	**47.17**	**51.03**	**42.23**	16.00	**21.36**	**21.98**	**37.64**	**29.00**	**104.85**
CVS (3B)	42.61	55.96	38.41	62.77	1104.09	41.93	49.31	41.00	**16.36**	17.68	17.88	36.35	28.60	98.51

Table 14: Detailed results on Vision-Flan for robustness to the target model (Qwen2-VL-2B). With the evaluator fixed to Qwen2-VL-7B-Instruct, we replace the target model from LLaVA-1.5-7B to Qwen2-VL-2B to evaluate the robustness of CVS under target model changes. Experiments are conducted using data subsets of 5%, 10%, and 15%. The best result in each column is highlighted in bold.

(a) Detailed results of CVS robustness to target model (Qwen2-VL-2B) changes using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	51.84	34.21	72.03	74.96	1294.90	61.68	68.99	68.94	52.00	81.86	50.67	42.58	74.80	100.00
Random	**52.40**	40.72	**71.88**	**75.75**	**1430.27**	53.35	58.85	**67.13**	48.20	76.36	49.41	**40.58**	74.10	98.34
CVS	48.86	**49.06**	71.50	74.96	1269.53	**55.33**	**61.25**	67.10	**49.12**	**79.16**	**51.58**	39.20	**74.50**	**99.65**

(b) Detailed results of CVS robustness to target model (Qwen2-VL-2B) changes using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	51.84	34.21	72.03	74.96	1294.90	61.68	68.99	68.94	52.00	81.86	50.67	42.58	74.80	100.00
Random	**52.09**	35.91	72.60	75.16	**1398.61**	55.58	62.03	67.88	**50.96**	78.10	51.61	40.36	74.30	98.64
CVS	48.37	**44.26**	**73.57**	**75.36**	1241.04	**57.47**	**62.89**	**68.07**	50.48	**80.85**	**53.02**	**40.91**	**75.10**	**100.09**

(c) Detailed results of CVS robustness to target model (Qwen2-VL-2B) changes using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	51.84	34.21	72.03	74.96	1294.90	61.68	68.99	68.94	52.00	81.86	50.67	42.58	74.80	100.00
Random	49.86	**39.80**	71.77	74.71	1194.15	**60.30**	**66.15**	**68.55**	44.44	77.15	51.18	40.06	74.40	97.80
CVS	**52.08**	36.21	**74.27**	**75.16**	**1343.07**	57.22	62.72	68.30	**53.04**	**80.69**	**51.51**	**41.30**	**74.90**	**99.65**

Table 15: Detailed results on Vision-Flan for robustness to the target model (LLaVA-1.5-13B). With the evaluator fixed to Qwen2-VL-7B-Instruct, we replace the target model from LLaVA-1.5-7B to LLaVA-1.5-13B to evaluate the robustness of CVS under target model changes. Experiments are conducted using data subsets of 5%, 10%, and 15%. The best result in each column is highlighted in bold.

(a) Detailed results of CVS robustness to target model (LLaVA-1.5-13B) changes using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	49.02	55.92	35.82	70.60	1318.89	56.44	60.74	57.67	19.44	20.03	19.87	35.65	29.50	100.00
Random	**43.02**	**56.82**	34.35	**63.46**	**1178.91**	36.25	41.23	**45.66**	14.88	14.87	18.17	32.17	**29.00**	85.12
CVS	39.54	56.80	**36.64**	63.21	920.15	**39.35**	**45.37**	43.65	**18.12**	**18.24**	**21.95**	**33.50**	28.40	**88.34**

(b) Detailed results of CVS robustness to target model (LLaVA-1.5-13B) changes using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	49.02	55.92	35.82	70.60	1318.89	56.44	60.74	57.67	19.44	20.03	19.87	35.65	29.50	100.00
Random	**44.98**	**58.13**	36.43	**68.26**	**1176.93**	**47.60**	**55.56**	**49.63**	16.56	16.97	19.08	31.35	29.60	92.26
CVS	41.27	57.48	**37.49**	66.58	1038.90	43.04	48.45	47.80	**17.60**	**19.03**	**21.73**	**35.66**	**29.80**	**92.28**

(c) Detailed results of CVS robustness to target model (LLaVA-1.5-13B) changes using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	49.02	55.92	35.82	70.60	1318.89	56.44	60.74	57.67	19.44	20.03	19.87	35.65	29.50	100.00
Random	**45.25**	**56.96**	**38.03**	65.05	**1176.15**	42.70	46.99	**48.32**	**16.60**	18.38	18.23	**34.86**	29.40	91.14
CVS	42.54	56.55	37.02	**67.18**	1150.28	**44.07**	**51.63**	46.11	15.80	**19.07**	**21.24**	32.30	**30.50**	**91.85**

Table 16: Detailed results on Vision-Flan for visual anchoring. To evaluate the necessity of the visual anchoring term in CVS, we remove the image information I from the conditional probability $P(Y \mid I, A)$ in the denominator and compare the resulting variant with random sampling using data subsets of 5%, 10%, and 15%. The best result in each column is highlighted in bold.

(a) Detailed results of the visual anchoring ablation study using a 5% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**42.93**	55.79	**36.54**	**60.34**	**1098.02**	**36.40**	**26.19**	37.82	**15.96**	**16.46**	**19.99**	**34.77**	27.80	**92.96**
CVS	36.12	**56.29**	33.08	57.56	493.39	31.19	14.69	**37.89**	14.92	15.99	17.29	33.08	**28.10**	82.44

(b) Detailed results of the visual anchoring ablation study using a 10% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**43.31**	**56.02**	**36.13**	60.92	**1152.77**	**39.61**	31.88	**44.50**	**16.52**	17.27	18.08	**34.36**	27.90	**95.30**
CVS	35.03	52.96	33.35	**64.06**	501.64	36.17	**41.75**	41.22	14.48	**18.15**	**19.80**	30.84	**28.70**	89.48

(c) Detailed results of the visual anchoring ablation study using a 15% data subset.

Method	GQA	VizWiz	TextVQA	SQA-I	MME	MMB-CN	MMB-EN	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	OCRBench	ARP
Full Data	47.30	55.88	35.83	61.43	1270.40	50.26	55.67	52.14	15.68	15.80	15.53	35.32	26.20	100.00
Random	**43.19**	**56.33**	**35.41**	59.59	**1230.04**	**42.44**	38.06	41.06	**15.52**	16.26	16.26	**32.27**	**27.60**	**93.82**
CVS	35.02	53.42	30.81	**63.31**	499.41	41.15	**46.82**	**41.97**	14.28	**17.58**	**17.04**	31.42	27.30	88.44



Q: "Identify the breed of the dog in the image.\n<image>"
A: "The dog breed is a curly-coated retriever."



Q: "Here is an image of some art. I want to know what type of painting it is. Among others, some types could be: religious, self-portraits, or landscapes\n<image> "
A: "This falls into the 'genre' type."

Figure 6: Representative examples selected by the Low strategy.

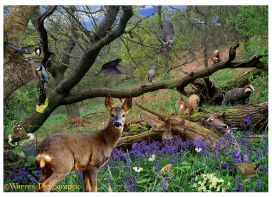


Q: "Here is a detailed conversation between a person and a robot. They are discussing some questions and answers about a picture. From the conversation, answer the question as the robot.\n\nHuman: anyone wearing glasses?\nRobot: no\nHuman: is this a girl's soccer team?\nRobot: mixed\nHuman: is it sunny in the picture?\nRobot: yes\nHuman: can you see the ball?\nRobot: yes\nHuman: can you see any adults?\nRobot: 1\nHuman: are the kids wearing uniforms?\nRobot: matching t shirts\nHuman: is this picture well lit?\nRobot: yes\nHuman: is this a professional photo?\nRobot: no\nHuman: is this a close up photo?\nRobot: no\nQuestion: what color are the matching shirts?\n<image>"
A: "Robot: blue team purple team."



Q: "<image>\nWhat is the next step? You are doing Cleaning Your Face and you have finished\n\nFigure out what type of face skin you have.\nWash your hands first.\nWash your face twice a day with lukewarm water and a gentle cleanser.\nFind out if exfoliation is right for your skin.\nRinse your face thoroughly after cleansing or exfoliating.\nPat your face dry with a clean, soft cloth.\nYou currently are at the step given by the image and the text \"Find out if exfoliation is right for your skin.\". The next step is "
A: "Rinse your face thoroughly after cleansing or exfoliating."

Figure 7: Representative examples selected by the High strategy.



Q: "What animal is in the foreground?\n<image>"
A: "A fox."



Q: "Is the train moving toward the camera or away from it?\n<image> "
A: "It appears to be moving away, since the headlights might be reflections from the platform lights."

Figure 8: Representative examples selected by the No strategy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide the claims and contributions at the end of the introduction, which have experimental support.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include theoretical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section 2.2, Section 3.1, Appendix C, Appendix I, and Appendix L for the algorithmic workflow, experimental setup, specific prompts, baseline hyperparameter settings, and detailed benchmark results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please see the anonymous GitHub link provided in the abstract, which includes the code and reproduction guidance for the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Section 2.2, Section 3.1, Appendix C, Appendix I for the algorithmic workflow, experimental setup, specific prompts, and baseline hyperparameter settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not report error bars, confidence intervals, or repeated-run statistical significance tests. First, CVS only relies on model inference, so its selection results are very stable and are largely unaffected by random seeds. Second, the results of CVS are based on two visual instruction tuning datasets, multiple data budgets, and 12 benchmarks, where its performance is clearly superior; therefore, it is unlikely that the conclusions are caused by random seed variation, and running multiple random seeds would be computationally prohibitive.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.1 states that the experiments are conducted on 4 NVIDIA RTX A6000 GPUs, and Table 3 reports the time required by CVS for data selection.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have carefully read the NeurIPS Code of Ethics and verified that our paper aligns with the requirements.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is foundational in nature and proposes a data selection method. It does not introduce a new deployed model, application, or decision-making system. Any potential societal effects are indirect and depend on how downstream models trained with the selected data are used. Thus, we do not anticipate direct societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new high-risk pretrained model, image generator, or scraped dataset. It releases a data selection method and experimental code, while relying on existing public datasets and VLLMs.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- This work proposes a data selection method and does not release a new high-risk model, image generator, or scraped dataset. Therefore, the safeguard requirements for controlled release are not applicable.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use publicly released datasets, models, and software packages, including Vision-Flan[48], The Cauldron[17], LLaVA[23], Qwen[4, 43], LLaVA-OneVision [3], InternVL3 [54], and vLLM[16], all of which are cited appropriately in the paper and used in accordance with their licenses or terms of use. Please see Section 3.1, Appendix H, and the anonymous GitHub repository for asset sources and usage details.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces the CVS implementation and associated experimental scripts as new assets. These assets are documented in the anonymous GitHub repository linked in the abstract, including usage instructions, reproduction guidance, and configuration details.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The use of vision-language large models is an essential component of the research, as they serve as scoring models, target models trained on the selected subsets, and the models evaluated in downstream benchmarks.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.