

An Embarrassingly Simple Data Selection Strategy for Million-Scale Multi-Modal Instruction-Tuning

Mengyao Lyu^{1,2}, Yan Li³, Huasong Zhong³, Hui Chen^{1,2},
Jungong Han^{1,2}, Guiguang Ding^{1,2,†}, Zhenheng Yang³

¹Tsinghua University, ²BNRist, ³Bytedance
mengyao.lyu@outlook.com dinggg@tsinghua.edu.cn

Abstract

Recent research demonstrates that pretrained large language models (LLMs) require minimal but carefully curated data for instruction-tuning. Building on LLMs, multi-modal models further contend with far larger data volumes, heterogeneous sources, and cross-modal alignment, amplifying both the necessity and the complexity of data selection. However, existing methods rely on **deeply extracted** embeddings, gradients or **narrow** quality scores that are time-consuming to compute and limited in expressiveness, frequently failing to surpass a simple random-sampling baseline. To overcome these limitations, we step back from complex embedding spaces and target the fundamental objective of instruction tuning: how and what knowledge should be elicited from the pretrained model and exposed to users. We leverage interaction Styles as a proxy for structural diversity, and decompose data quality into capability-specific Scores, yielding an **explicit** yet **Rich** (SSR) profile for each sample. We further cast the selection as submodular coverage maximization over the capability–style grid, providing an approximation guarantee. SSR consistently delivers superior performance across two distinct architecture stacks, LLaVA-1.5 (Vicuna + CLIP) and LLaVA-OneVision (Qwen2 + SigLIP), two large-scale data pools, six data budget settings, and 14 multi-modal benchmarks. To our knowledge, SSR is the first data selection method validated at million-scale, achieving 99.1% of full-data performance using only 30% of a 2.6M pool. Furthermore, the SSR profile models and selected subsets transfer training-free across data pools and model architectures with capability extensibility.

1 Introduction

Data matters for the successful scaling of large foundation models (Li et al., 2024a; Wettig et al.,

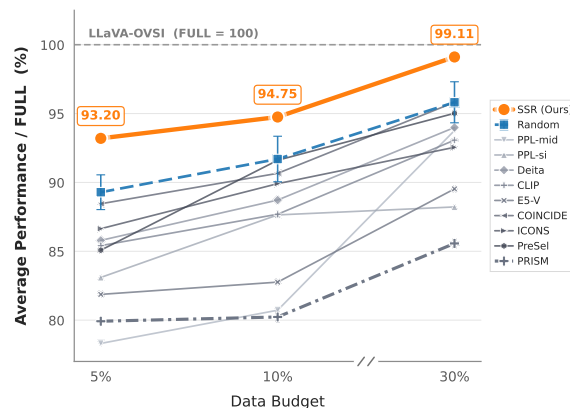


Figure 1: **Million-scale selection exposes the failure of prior methods.** Failing to beat random sampling across 5%, 10%, and 30% budgets on the 2.6M data pool, existing methods expose a scalability gap that SSR closes with consistent gains.

2024; Liu et al., 2024c; Lu et al., 2024a; Li et al., 2025), particularly during supervised fine-tuning (SFT), which aligns pre-trained models to follow diverse instructions across general and specialized domains. Early work on large language models (LLMs) focused on filtering million-scale SFT pools for deduplication (Lee et al., 2022; Elazar et al., 2024), quality, and safety (Joulin et al., 2016; Penedo et al., 2023; Dubey et al., 2024; Team et al., 2024; Chung et al., 2024). LIMA (Zhou et al., 2023) then argued that “less is more” for alignment: with 1,000 carefully curated samples, it showed that LLM knowledge is acquired during pre-training, so instruction tuning needs little data provided it is of high quality and diverse. This sparked a wave of automated selection methods (Lu et al., 2024a; Xia et al., 2024; Liu et al., 2024b). However, subsequent surveys (Diddee and Ippolito, 2025; Xia et al., 2025) reveal that these methods are highly sensitive to data budgets, sources, and evaluation protocols, and frequently fail to beat random sampling. They typically depend on *heavy extracted* representations such as high-dimensional embeddings or gradients (Liu et al., 2024c; Pang et al., 2025; Li et al., 2024b; Xia et al., 2024; Safaei et al., 2025), which are costly to scale and may hurt

[†]Corresponding author.

Project page: <https://lyumengyao.github.io/projects/ssr>.

transferability (Liu et al., 2024b). A complementary line rates data with proprietary LLMs (e.g., GPT-4) on heuristic metrics (Wettig et al., 2024; Liu et al., 2024c; Chen et al., 2024a), but the resulting criteria are typically too *narrow*, limiting expressiveness (Liu et al., 2024c; Pang et al., 2025) and generality (Wettig et al., 2024).

These failure modes are further exacerbated in Multi-modal LLMs (MLLMs): massive token volumes, heterogeneous sources, and diverse downstream requirements turn *million-scale* fine-tuning into a critical yet underexplored bottleneck for data selection (Fig. 1). *First*, *model-centric* statistics fail to scale at million-level MLLM data volumes. High-dimensional intrinsic-feature clustering (Lee et al., 2024; Safaei et al., 2025; Bi et al., 2025), gradient-based distribution matching (Lee et al., 2024), and loss re-weighting via score embeddings (Chen et al., 2024d) grow prohibitive in compute or brittle under distribution shift, leaving state-of-the-art (SOTA) automated selection research stuck at thousand-scale datasets and lagging orders of magnitude behind today’s million-sample MLLM pools. *Second*, selection signals anchored in model-specific latents do not readily transfer across settings by construction. Each shift in target architecture poses new extraction costs and adaptation challenges, and such signals remain brittle under changes in budget, source, or evaluation protocol (Diddee and Ippolito, 2025; Xia et al., 2025). *Third*, existing methods yield selections too narrow for the broad generality that MLLM demands. Beyond model-specific latents, scalar criteria collapse each sample into one or a few dimensions, leaving the diverse axes of downstream MLLM capabilities uncovered, while noise and variability in multi-modal curation (Chen et al., 2024b; Li et al., 2025; Liu et al., 2024a) further make selection for target-model generality fragile. Experiments in Sec. 4.2 validate these vulnerabilities and motivate a new paradigm for multi-modal selection.

Toward such a paradigm, we ground our method in the fundamental objective of SFT. Beyond “less is more”, LIMA’s Superficial Alignment Hypothesis (Zhou et al., 2023) posits that a model’s core knowledge is acquired during pre-training and fine-tuning mainly teaches how to articulate that knowledge. This motivates two axes for sample valuation: an interaction *style* encoding *how* the knowledge is delivered, and a capability *score* encoding *what* knowledge should be prioritized. Interaction Styles and capability Scores together form an ex-

plicit yet **Rich** profiling framework (**SSR**): each sample receives a style label (caption, yes/no) and a vector of transparent, human-interpretable capability scores, both judged per sample superficially without latent computation. We cast selection over this Cartesian-product coverage grid as a monotone submodular maximization under cardinality and partition-matroid constraints. This design departs from previous methods and directly addresses the challenges raised above. *First*, against model-centric statistics, SSR adopts a *data-centric* profile: each sample is judged explicitly and superficially without latent computation, which scales linearly to millions of samples. *Second*, against model-centric latents, SSR’s profile is intrinsic to each sample: style and capability are read superficially and explicitly from the sample itself, so different models and data pools converge on the same profile. The profile therefore transfers training-free across data domains and architectures, where latent-anchored methods drift. *Third*, against fragile generality, our capability–style coverage is rich, with $(1 - 1/e)$ coverage and EF1 fairness guarantees across cells (Thms. 1, 2). SSR is therefore scalable, transferable, and generalizable by construction.

We validate SSR on three architecture families, i.e., LLaVA-1.5 (Vicuna + CLIP) (Liu et al., 2024a), LLaVA-OneVision (Qwen2 + SigLIP) (Li et al., 2025), and the vision-centric Cambrian-1 (Llama-3 + a multi-encoder ensemble) (Tong et al., 2024), and three data pools at different scales: the 624K ShareGPT4V pool, the 2.6M LLaVA-OneVision pool, and the 7M Cambrian-7M pool. Despite its simplicity, SSR consistently outperforms baselines across six data budgets and two model sizes, benchmarked over 14 diverse evaluations, and transfers training-free to out-of-domain data pools and unseen architectures. Our main contributions are:

- We propose a SFT-aligned selection paradigm, SSR, that delivers key multi-modal capabilities through diverse interactional styles, yielding an explicit, model-agnostic profiling of multi-modal instruction data.
- We are the first to scale principled multi-modal SFT data selection to the million-scale regime, an urgent step as today’s MLLM training pools have outgrown the thousand-scale convention of prior selection research.
- We achieve SOTA effectiveness, transferability and scalability across 3 architecture families, 6 data budgets, and 14 benchmarks.

2 Related Work

We draw on data efficiency studies in LLMs, MLLMs, VLMs, and active learning, broadly categorized into hand-crafted heuristics, model-based indicators, and judge-based scores, which can be interchangeably or complementarily applied.

Hand-Crafted Heuristics use expert-defined quality criteria, e.g., relevance, clarity, diversity, safety, vocabulary, N-grams, length, for filtering (Team et al., 2023; Touvron et al., 2023; Dubey et al., 2024; Qin et al., 2024; Penedo et al., 2023). They are interpretable but labor-intensive, bias-prone, narrow, and potentially ill-suited to multi-modality.

Model-Based Indicators leverage target-model gradients, predictive distributions, or embeddings to assess uncertainty, learnability, and similarity (Evans et al., 2024; Liu et al., 2024f; Lyu et al., 2023, 2025; Sener and Savarese, 2018; Liu et al., 2024c,b; Settles, 2009), with strong in-domain results but limited interpretability and transferability (Diddee and Ippolito, 2025; Munjal et al., 2022). Proxy-model variants train classifiers (Joulin et al., 2016; Brown et al., 2020; Gao et al., 2025; Li et al., 2024a) or use proxy representations: COINCIDE (Lee et al., 2024) trains a 2B proxy on the target pool for coreset (Sener and Savarese, 2018) selection via high-dimensional clustering, ICONS (Wu et al., 2024) estimates sample influence via gradient similarity to a validation set, PreSel (Safaei et al., 2025) trains a reference model to compute modality relevance and task importance, and PRISM (Bi et al., 2025) scores samples by their correlation with a corpus-mean-centered pool of MLLM features: all anchor sample value in one model’s latents, and incur substantial overhead at scale. VLM-side multi-modal quality metrics target alignment (Maini et al., 2024; Gadre et al., 2023; Goyal et al., 2024) but correlate weakly with instruction-following and lead to redundancy.

Judge-Based Scoring (Liu et al., 2025, 2024c; Chen et al., 2024a; Pang et al., 2025) employs an external judge, often a proprietary API, as a cost-effective alternative to human annotation for candidate valuation. The resulting LLM-side quality dimensions are often too narrow to capture multi-modal variability, and additional sample-level or pairwise scoring with similarity thresholds (Liu et al., 2024c) does not scale globally, leaving judge-based selection largely unexplored for MLLMs. SSR addresses both: a multi-dimensional capability axis broadens beyond singular scalars, and an

interaction-style axis with submodular coverage replaces latent compute, making judge-based selection viable and effective at MLLM scale.

Offline vs. Online Selection. The methods above perform *offline* selection: a subset is chosen once from a static pool before fine-tuning, so the output is a reusable, auditable artifact independent of any particular training run. A complementary line performs *online* selection, where per-batch decisions track the state of the model being trained: OASIS (Lee et al., 2026) adaptively admits informative samples during continual visual instruction tuning, and PROGRESS (Yang et al., 2026) clusters the pool offline into concepts, then prioritizes them online by the model’s relative error-driven learning progress. Online selection reacts to what the model has learned, but couples selection to training, yields no transferable subset, and presumes a fixed target architecture. Offline selection instead amortizes profiling once and, in SSR’s data-centric form, transfers across pools and architectures. The two paradigms are complementary rather than competing, as PROGRESS’s illustrates; as an offline selector, SSR is compared against offline methods.

3 Method

Problem setup. Given a large-scale multi-modal instruction pool $D = \{X_1, \dots, X_N\}$ with $X_i = (I_i, Q_i, A_i)$ and a budget percentage $b\%$ ($b \in (0, 100]$, e.g., $b = 10$ or 30 in our experiments), we seek a subset $\hat{D} \subseteq D$ with $|\hat{D}| \leq b\% \cdot |D|$ that, when used to fine-tune a pretrained MLLM $\mathcal{M}_\theta$, maximizes downstream instruction-following ability. Inspired by the Superficial Alignment Hypothesis (SAH) (Zhou et al., 2023), we ground selection in two complementary axes of sample value: a capability score that encodes *what* knowledge each sample conveys, and an interaction style that encodes *how* that knowledge is delivered. Let $C = \{c_1, \dots, c_{|C|}\}$ be the set of capabilities and $\Sigma = \{\sigma_1, \dots, \sigma_{|\Sigma|}\}$ the set of interaction styles (taxonomy derivation in Supp. B.4). The Cartesian product $C \times \Sigma$ defines $|C| \cdot |\Sigma|$ capability–style *cells*; covering this grid is the formal target of SSR.

Pipeline. As illustrated in Fig. 2, SSR comprises two algorithmic stages, accompanied by inherent benefits. (i) *Profiling*: assign every $X_i \in D$ a non-negative capability vector $r_i \in [0, 5]^{|C|}$ and an interaction-style indicator $g_i \in \{0, 1\}^{|\Sigma|}$ (Sec. 3.1). (ii) *Selection*: maximize a monotone submodular coverage objective F over the capability–style

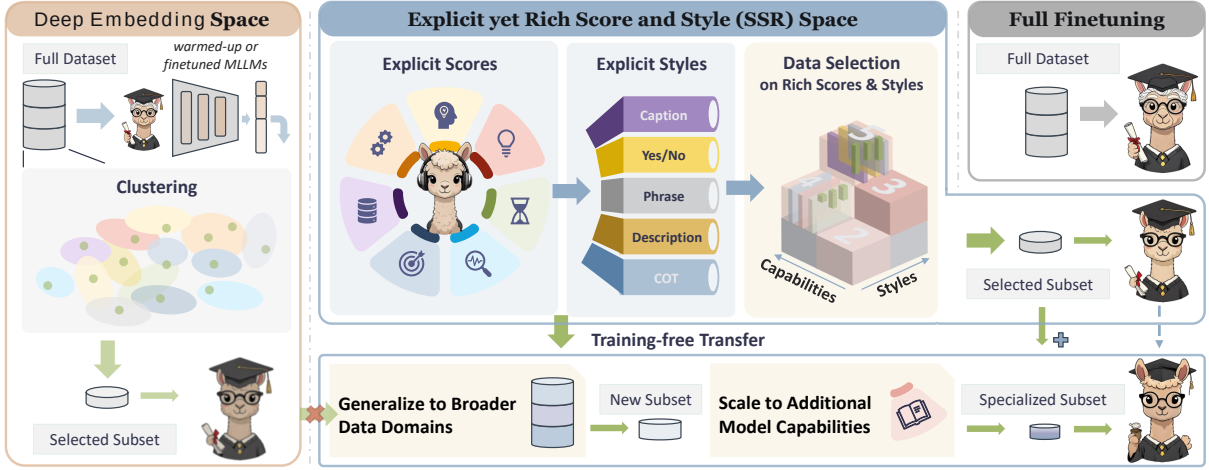


Figure 2: **Overview of the SSR framework compared to existing paradigms.** Full fine-tuning indiscriminately processes all data (Top Right), while previous selection methods (Left) rely on *model-centric* signals (e.g., latent embeddings). In contrast, our proposed SSR (Middle) adopts a *data-centric* profile: rich and explicit capability scores and an interaction style, which together drive submodular coverage maximization over the capability–style grid, yielding an efficient subset with superior performance, training-free transfer across data pools and target models, and seamless capability extensibility.

grid by stratified round-robin, which admits both a $(1 - 1/e)$ approximation bound and an envy-free-up-to-one-item (EF1) fairness guarantee (Lipton et al., 2004) (Sec. 3.2, 3.3). (iii) *Extensibility*: by carrying semantic rather than model-specific signals, the profile transfers across data pools and target architectures (Sec. 4.4), and the capability set C accommodates new skills (Supp. D).

3.1 SSR Profiling

The dominant paradigm for MLLM data selection is *model-centric*: per-instance value is computed from the target MLLM itself via gradient norms, influence functions, latent embeddings or predicted loss (Xia et al., 2024; Wu et al., 2024; Lee et al., 2024; Safaei et al., 2025). These signals are coupled to one model checkpoint, demand gradient or hidden-state access, and yield representations that offer limited interpretability and tend to transfer less readily across target models or data pools. In contrast, a *data-centric* metric assigns each sample a scalar quality score $s(X_i) \in \mathbb{R}$ by prompting a strong judge (Pang et al., 2025; Wettig et al., 2024; Chen et al., 2024a; Liu et al., 2025). Such scores are architecture-agnostic and semantically interpretable, but narrow in granularity.

Formally, for any summary s of a sample X , the data-processing inequality bounds the mutual information with the downstream SFT utility U as $I(s(X); U) \leq I(X; U)$; how much of $I(X; U)$ is recovered by s depends on its structure. One or a few broad criteria such as quality and complexity compress a sample’s value into too few axes to track the heterogeneity of MLLM SFT utilities. This gap is empirically witnessed by score-

based methods (Liu et al., 2024c; Pang et al., 2025; Liu et al., 2025) that re-inject diversity through an explicit post-hoc feature-dissimilarity-aware algorithm, a mitigation that re-introduces poor scalability and model-specific dependence. We instead read the axes of value off the structure of SFT to factorize a sample’s contribution into knowledge content and interaction style, two semantically distinct axes forming a *rich* per-sample profile:

- *Capability scores* $r_i \in [0, 5]^{|C|}$ rate *what* knowledge each instance contributes to each fine-grained capability (e.g., spatial reasoning, STEM knowledge, OCR). Scores are produced against a unified 0–5 rubric so they are commensurable across capabilities and samples.
- *Interaction styles* $g_i \in \{0, 1\}^{|\Sigma|}$ record *how* that knowledge is delivered, capturing the dominant interaction format (e.g., multiple choice, captioning) emphasized by LIMA’s SAH.

C and Σ are derived from open-source instruction data through a one-time procedure of discovery, expansion, and iterative refinement (Supp. B.4). The resulting $|C|$ capabilities and $|\Sigma|$ interaction styles, being explicitly semantic rather than model-specific, are interpretable and transferable.

The two axes combine into a non-negative tensor $q \in \mathbb{R}_{\geq 0}^{N \times |C| \times |\Sigma|}$ whose per-sample slice $q_i \in \mathbb{R}_{\geq 0}^{|C| \times |\Sigma|}$ is the rank-one outer product

$$q_{c\sigma}(i) = r_i^c \cdot g_i^\sigma, \quad (c, \sigma) \in C \times \Sigma. \quad (1)$$

Equivalently, $q_i = r_i \otimes g_i$ replicates sample i ’s capability scores along its active style channels ($g_i^\sigma > 0$) and vanishes on inactive cells.

3.2 Submodular Coverage Objective

Given target capabilities $\hat{C} \subseteq C$ and styles $\hat{\Sigma} \subseteq \Sigma$, we want $\hat{D}$ to cover the grid $\hat{C} \times \hat{\Sigma}$ broadly, not to greedily exhaust any single cell. We instantiate this intuition with the saturated coverage family of Lin and Bilmes (Lin and Bilmes, 2011): for each cell (c, σ) , the contribution saturates once a fraction of pool informativeness has been collected, encoding diminishing returns. Define

$$F(S) = \sum_{(c, \sigma) \in \hat{C} \times \hat{\Sigma}} w_{c\sigma} \min\left\{\tau_{c\sigma}, \sum_{i \in S} q_{c\sigma}(i)\right\}, \quad (2)$$

with cell weights $w_{c\sigma} \geq 0$ (uniform by default) and saturation thresholds $\tau_{c\sigma} > 0$ (defaulting to a fraction of pool capacity, $\tau_{c\sigma} = \alpha \cdot \sum_{i \in D} q_{c\sigma}(i)$ with $\alpha \in (0, 1]$). The min operator turns per-cell sums from modular into submodular: once a cell has accumulated $\tau_{c\sigma}$ worth of evidence, additional items in it contribute zero marginal value. Note that $w_{c\sigma}$ and α are analysis variables, not hyperparameters.

Proposition 1 (Submodularity). *The set function $F: 2^D \rightarrow \mathbb{R}$ defined by Eq. 2 is normalized, monotone non-decreasing, and submodular.*

Proof sketch. Each cell term composes a non-decreasing concave map $x \mapsto w_{c\sigma} \min\{\tau_{c\sigma}, x\}$ with the modular sum $\sum_{i \in S} q_{c\sigma}(i) \geq 0$, which is monotone submodular (Lin and Bilmes, 2011); F is their non-negative linear combination. Please refer to Supp. A.2 for full proof.

Structural contrast with scalar top- k . Top- k selection on any scalar quality score $s(X_i)$ maximizes a *modular* set function $\sum_{i \in S} s(X_i)$ under a cardinality constraint: the optimum is exact, but the objective is additive in items and therefore does not penalize redundancy among them. Eq. 2 is submodular by construction (Prop. 1): the saturation operator encodes diminishing returns within each cell, so once a cell is well covered, additional items in it contribute less to F than items in under-covered cells. This change in problem class, from modular to submodular under a partition matroid, is what enables the approximation and fairness guarantees we exploit in Sec. 3.3; the same guarantees do not arise for scalar scoring (Liu et al., 2024c; Chen et al., 2024a) or single-axis tagging (Lu et al., 2024a), which lack a coverage structure to optimize.

3.3 Approximation and Guarantees

Exact maximization of Eq. 2 under a cardinality or per-cell budget is NP-hard in general, subsuming maximum coverage. We approximate it with a

stratified round-robin selector that admits both a $(1 - 1/e)$ approximation bound and an envy-free-up-to-one-item (EF1) allocation guarantee across capability–style cells.

Algorithm. For each $(c, \sigma) \in \hat{C} \times \hat{\Sigma}$, form the candidate group $G_{c\sigma} = \{X_i \in D : q_{c\sigma}(i) > 0\}$ and sort it in descending order of $q_{c\sigma}(\cdot)$, ties broken by a fixed rule. Set the per-cell budget $k_{c\sigma} = \lfloor b\% \cdot |D| / (|\hat{C}| |\hat{\Sigma}|) \rfloor$, fix any cyclic ordering of cells, and initialize per-cell claim lists $A_{c\sigma} \leftarrow \emptyset$ and the selected pool $\hat{D} \leftarrow \emptyset$. For each round $t = 1, 2, \dots$ and each cell (c, σ) in the cyclic order: if $|A_{c\sigma}| < k_{c\sigma}$ and $G_{c\sigma} \setminus \hat{D} \neq \emptyset$, let i^* be the highest- $q_{c\sigma}$ element of $G_{c\sigma} \setminus \hat{D}$ and set $A_{c\sigma} \leftarrow A_{c\sigma} \cup \{i^*\}$ and $\hat{D} \leftarrow \hat{D} \cup \{i^*\}$. Terminate once every cell has $|A_{c\sigma}| = k_{c\sigma}$ or $G_{c\sigma} \subseteq \hat{D}$. A sample may appear in multiple candidate groups, but once selected it leaves all remaining queues, so each sample is claimed by exactly one cell and $|\hat{D}| = \sum_{(c, \sigma)} |A_{c\sigma}| \leq \sum_{(c, \sigma)} k_{c\sigma} \leq b\% \cdot |D|$, with strict inequality only when a queue exhausts early.

Guarantees. Two matching theoretical bounds hold for the round-robin output; formal statements, the lifted ground-set construction they rest on, and full proofs are deferred to Supp. A.

- *(1 - 1/e) coverage approximation* (Thm. 1). Lifting the per-cell budgets to a partition matroid yields a coverage objective $F\hat{F}$ that decomposes additively across cells; an idealized non-competing round-robin coincides with discrete greedy run independently inside each block, so the per-block Nemhauser–Wolsey–Fisher bound (Nemhauser et al., 1978) delivers a $(1 - 1/e)$ -factor approximation to the per-cell-budgeted optimum, which the implemented selection inherits by containment (Lem. 1).
- *EF1 fairness across cells* (Thm. 2). Treating each cell (c, σ) as an additive agent with valuation $v_{c\sigma}(S) = \sum_{i \in S} q_{c\sigma}(i)$, the allocation $\{A_{c\sigma}\}$ produced by round-robin is envy-free up to one item (Lipton et al., 2004): no cell can be starved by more than a single item’s value relative to any other. This rules out scalar top- k ’s failure mode of pouring the entire budget into one cell.

3.4 Profile Sources

The selection machinery in Sec. 3.2–3.3 consumes only the cell-conditioned value tensor q . Any procedure that produces r_i and g_i for every $X_i \in D$ is a valid profile source. We outline two complementary instantiations; both are validated in Sec. 4.3.

	MMBench _{en-v1.1}	MMStar	MMMU	MMVet	BLINK	MMT-Bench	MME	AI2D	ScienceQA	MathVista _{mini}	>Rand	/FULL
Budget: 5%												
Random	73.74	47.98	43.70	42.34	50.61	58.87	2004.50	73.07	81.52	45.47	-	89.29
PPL-mid	67.34	45.27	38.98	30.18	45.27	54.33	1887.71	66.74	74.76	31.40	0/10	78.31
PPL-si	71.98	44.67	38.48	35.14	54.10	57.98	1856.79	67.84	78.24	36.50	1/10	83.10
Deita _{top-k}	72.91	47.47	41.28	40.23	<u>52.59</u>	56.57	1956.50	70.76	79.57	36.10	1/10	85.79
CLIP	<u>74.23</u>	47.27	40.08	35.73	<u>52.96</u>	56.73	1902.65	<u>73.61</u>	78.63	39.80	3/10	85.41
E5-V	70.90	43.00	38.78	38.44	49.94	54.65	1810.47	66.58	77.54	37.40	0/10	81.87
COINCIDE	72.76	<u>48.33</u>	43.17	45.60	49.43	57.53	1852.66	<u>73.15</u>	79.62	45.40	3/10	88.45
ICONS	66.72	<u>52.20</u>	41.18	38.03	47.92	55.96	1811.13	<u>76.20</u>	<u>83.64</u>	<u>46.90</u>	4/10	86.64
PreSel	72.68	<u>48.33</u>	40.96	35.18	<u>52.21</u>	55.68	1954.35	69.01	79.38	40.70	2/10	85.08
PRISM	65.56	44.07	41.07	33.49	45.90	51.84	1807.73	65.54	76.60	40.30	0/10	79.92
SSR (Ours)	77.79	53.33	43.27	<u>43.53</u>	<u>51.83</u>	59.16	1938.68	77.66	88.45	52.00	8/10	93.20
Budget: 10%												
Random	74.57	51.57	44.72	42.91	52.59	58.99	2033.28	74.42	84.33	47.80	-	91.70
PPL-mid	63.54	46.87	39.08	36.93	45.90	54.30	1831.03	67.23	73.87	39.50	0/10	80.72
PPL-si	<u>74.69</u>	49.80	41.28	40.60	53.09	57.95	1841.11	<u>75.16</u>	80.71	40.40	3/10	87.64
Deita _{top-k}	<u>75.39</u>	48.80	43.77	42.25	<u>54.48</u>	57.40	1996.34	71.60	78.33	40.80	2/10	88.72
CLIP	<u>75.23</u>	49.87	40.38	37.16	<u>53.59</u>	<u>59.35</u>	1921.04	<u>76.62</u>	80.07	41.00	4/10	87.69
E5-V	70.51	45.13	38.78	39.59	50.57	55.10	1787.94	68.94	77.54	37.20	0/10	82.76
COINCIDE	<u>75.23</u>	49.73	<u>44.77</u>	42.52	50.69	58.71	2027.58	<u>74.77</u>	82.05	47.00	3/10	90.66
ICONS	71.67	<u>53.33</u>	44.17	40.46	49.18	57.40	1789.60	<u>76.65</u>	<u>85.23</u>	<u>51.10</u>	4/10	89.91
PreSel	<u>76.93</u>	53.87	42.82	33.94	55.23	57.28	1904.98	79.44	92.86	<u>50.40</u>	6/10	91.61
PRISM	64.63	<u>45.47</u>	41.31	30.78	45.65	53.21	1849.71	66.13	76.20	42.10	0/10	80.23
SSR (Ours)	77.32	<u>53.27</u>	45.06	42.98	54.10	59.61	2045.00	<u>78.76</u>	<u>89.94</u>	52.40	10/10	94.75
Budget: 30%												
Random	78.25	54.60	44.40	46.10	55.23	59.61	2092.60	78.28	88.32	52.57	-	95.82
PPL-mid	73.99	<u>54.93</u>	43.97	41.01	53.09	58.78	2036.54	77.20	87.01	<u>56.40</u>	2/10	93.77
PPL-si	72.52	48.33	42.57	43.62	51.83	55.07	1976.46	76.55	78.48	42.20	0/10	88.22
Deita _{top-k}	76.93	54.13	43.67	44.04	55.11	59.66	2042.63	79.50	83.54	50.30	2/10	93.99
CLIP	74.30	53.80	43.07	45.87	51.95	<u>59.16</u>	2039.14	<u>80.02</u>	83.99	48.80	1/10	93.07
E5-V	74.30	46.07	43.27	47.80	50.32	57.85	1955.13	74.45	81.61	43.70	1/10	89.53
COINCIDE	78.02	<u>55.47</u>	45.66	<u>46.24</u>	52.84	<u>59.80</u>	2047.37	<u>79.73</u>	84.33	<u>55.10</u>	6/10	95.82
ICONS	71.90	53.40	43.87	42.25	50.32	59.23	1985.64	78.21	86.76	<u>54.10</u>	1/10	92.55
PreSel	76.86	<u>57.20</u>	41.89	36.93	59.14	58.91	1971.60	81.61	93.75	<u>55.20</u>	5/10	95.02
PRISM	64.32	49.60	41.77	40.60	46.66	54.01	1898.22	70.43	81.26	47.40	0/10	85.57
SSR (Ours)	79.57	57.53	<u>44.87</u>	48.49	<u>56.24</u>	59.83	2132.93	<u>81.25</u>	<u>92.46</u>	57.40	10/10	99.11
FULL												
LLaVA-OVSI	80.57	59.40	45.16	47.16	56.87	60.73	2117.56	81.87	92.76	59.60	-	100

Table 1: **Million-scale scalability and benchmark-wide generality.** Performance on multi-modal benchmarks across 5%, 10% and 30% budgets of LLaVA-OVSI; best result in **boldface**, underline if a method beats the random baseline. The >Rand column reports the number of benchmarks where the method outperforms random sampling; /FULL compares to the FULL dataset. Among all evaluated methods, SSR is the only one that leads the >Rand metric at every budget, attaining a clean 10/10 sweep at the 10% and 30% budgets and recovering 99.1% of full-data performance with 30% data.

Direct judge query (no finetuning). A capable judge (a proprietary API such as GPT-4o, an open-weight MLLM such as Qwen2-VL, or an in-house fine-tuned MLLM checkpoint) is prompted with a fixed rubric to emit r_i and g_i for every $X_i \in D$. This path requires no training; its cost scales linearly in $|D|$, with per-call price or local inference compute as the constant factor.

Distilled judge (optional, separately trained). When the judge cost on the full pool D exceeds the available annotation budget, we amortize annotations by fine-tuning two *independent* student MLLMs on a small judge-labeled subset $D_J \subset D$, following the standard visual-instruction-tuning recipe (Li et al., 2025) under the next-token-prediction objective

$$\mathcal{L}_i = - \sum_t \log p_\theta(y_{i,t}^J | y_{i,<t}^J, X_i), \quad (3)$$

where y_i^J is the score string and the style string. At

inference, $\hat{r}_i$ and $\hat{g}_i$ are combined via Eq. 1 into q . We query GPT-4o on $|D_J|/|D| \approx 15\%$ data for distillation, and validate this choice on Supp. C.

4 Experiments

Sec. 4.2 reports million-scale selection on the 2.6M multi-modal data pool, where, among all evaluated methods, SSR is the only one that scales and delivers consistent benchmark-wide gains over prior baselines. Sec. 4.3 dissects the richness mechanism behind these gains and probes robustness to annotator choices. Sec. 4.4 demonstrates training-free transfer of SSR across architectures and data pools, including zero-shot transfer to a third architecture family on a 7M pool. Additional details (Supp. B), SSR model fidelity (Supp. C), capability scaling (Supp. D), results in 6 budget settings (Supp. D), and efficiency (Supp. E) are deferred to appendix.

R _{scr} R _{sty}		MMBench _{en-v1.1}	MMStar	MMMU	MMVet	BLINK	MMT-Bench	MME	AI2D	ScienceQA	MathVista _{mini}	>Rand /FULL
Budget: 5%												
Random		73.74	47.98	43.70	42.34	50.61	58.87	2004.50	73.07	81.52	45.47	- 89.29
SP _{GPT-4o}		<u>75.85</u>	<u>51.27</u>	42.97	44.27	51.95	58.14	1940.27	<u>73.61</u>	81.46	45.00	5/10 90.14
ScrR _{GPT-4o}	✓	<u>74.23</u>	<u>49.07</u>	43.07	41.56	<u>52.08</u>	57.56	1996.27	<u>74.77</u>	80.12	44.80	4/10 89.17
StyR _{GPT-4o}	✓	<u>76.48</u>	<u>52.52</u>	43.85	41.58	<u>52.46</u>	58.89	1945.68	<u>76.64</u>	86.88	49.87	8/10 92.07
SSR _{Qwen2-VL}	✓	<u>75.08</u>	<u>51.00</u>	45.16	42.57	<u>52.71</u>	57.37	1955.78	<u>74.74</u>	<u>84.88</u>	<u>48.90</u>	8/10 91.37
SSR _{LLaVA-OVSI}	✓	<u>77.40</u>	<u>50.60</u>	44.77	41.10	54.35	58.62	1952.97	<u>75.81</u>	87.75	40.40	6/10 90.68
SSR _{GPT-4o}	✓	77.79	53.33	43.27	43.53	51.83	59.16	1938.68	77.66	88.45	52.00	8/10 93.20
Budget: 10%												
Random		74.57	51.57	44.72	42.91	52.59	58.99	2033.28	74.42	84.33	47.80	- 91.70
SP _{GPT-4o}		<u>77.24</u>	50.40	44.27	42.52	<u>53.47</u>	<u>59.48</u>	2084.39	<u>76.07</u>	81.36	46.10	5/10 91.73
ScrR _{GPT-4o}	✓	<u>73.76</u>	49.40	44.77	42.80	46.91	57.24	2000.17	<u>75.39</u>	83.79	44.40	2/10 89.27
StyR _{GPT-4o}	✓	<u>77.72</u>	<u>54.36</u>	44.17	<u>44.62</u>	<u>54.05</u>	59.60	1928.81	<u>78.66</u>	89.75	52.80	8/10 94.61
SSR _{Qwen2-VL}	✓	<u>76.24</u>	<u>53.33</u>	44.87	45.60	55.11	<u>59.16</u>	2012.94	<u>76.75</u>	<u>87.11</u>	<u>52.70</u>	9/10 94.59
SSR _{LLaVA-OVSI}	✓	77.79	54.40	44.67	42.02	54.98	58.23	2013.74	78.85	89.59	42.00	5/10 92.72
SSR _{GPT-4o}	✓	<u>77.32</u>	<u>53.27</u>	45.06	42.98	54.10	59.61	2045.00	<u>78.76</u>	89.94	52.40	10/10 94.75
Budget: 30%												
Random		78.25	54.60	44.40	46.10	55.23	59.61	2092.60	78.28	88.32	52.57	- 95.82
SP _{GPT-4o}		<u>77.86</u>	53.13	45.76	48.03	54.85	58.78	2050.69	<u>78.92</u>	86.91	<u>55.80</u>	4/10 96.31
ScrR _{GPT-4o}	✓	<u>77.09</u>	52.67	43.47	44.31	53.59	58.23	2024.57	<u>79.11</u>	87.90	52.20	1/10 93.93
StyR _{GPT-4o}	✓	<u>78.27</u>	<u>55.84</u>	42.87	43.11	54.43	59.44	2079.25	<u>80.42</u>	<u>92.15</u>	<u>55.96</u>	5/10 96.07
SSR _{Qwen2-VL}	✓	<u>78.02</u>	<u>57.13</u>	43.07	47.39	<u>55.49</u>	60.89	2096.60	81.64	<u>90.28</u>	57.40	8/10 97.91
SSR _{LLaVA-OVSI}	✓	<u>77.55</u>	54.53	43.37	44.72	55.23	58.59	1980.48	81.02	91.87	49.60	2/10 94.73
SSR _{GPT-4o}	✓	79.57	57.53	44.87	48.49	56.24	<u>59.83</u>	2132.93	<u>81.25</u>	92.46	57.40	10/10 99.11

Table 2: **Richness drives generality, robustly across annotator sources.** Both rich scores (ScrR) and styles (StyR) are necessary for accuracy and generality: dropping either degrades coverage, and the single-criterion, i.e., *quality* SP(oor) baseline fails to cover the diverse axes of evaluation. {method}_{source} indicates querying scores and styles from {source} for {method}. With either distilled proxies (SSR_{GPT-4o}) or direct open-weight model prompting (SSR_{Qwen2-VL}, SSR_{LLaVA-OVSI}), SSR consistently beats prior baselines, confirming the efficacy stems from the explicit, rich SSR formulation that is robust to annotators.

4.1 Experimental Setup

Data pool. We base our main experiments on the single-image SFT stage of LLaVA-OneVision (LLaVA-OVSI) (Li et al., 2025), the open-source, open-data SOTA MLLM series. Within its 3.2M high-quality instances¹, 2.6M multi-modal data are openly available, which we consider as the full data pool and perform sample selection on it. This well-curated dataset covers over 90 sources, encompassing natural images, math and reasoning questions, documents, charts, screenshots *etc.* For our transfer experiments (Sec. 4.4), the earlier ShareGPT4V (Chen et al., 2024b) and the Cambrian-7M pool (Tong et al., 2024) are adopted.

Target model. SSR is evaluated in three distinct architectures, LLaVA-OneVision (Qwen2 + SigLIP), LLaVA-1.5 (Vicuna + CLIP), and Cambrian-1 (Llama-3 + a four-encoder ensemble).

Training setup. We utilize the stage-1.5-7B checkpoint of LLaVA-OneVision as the pretrained model for both SSR finetuning and single-image instruction tuning. To mitigate computational overhead during our extensive comparative experiments, we limit the maximum token length to 12k, which allows all training runs to be executed on 64 Nvidia H100 GPUs with a global batch size of 128. Be-

yond this modification, we strictly adhere to the hyperparameter settings and training protocols established in the official implementation. For our transfer experiments (Sec. 4.4), we adopt the LLaVA-1.5-7B architecture (Vicuna + CLIP) (Liu et al., 2024a) as the base model. Likewise, consistency is maintained by ensuring the fine-tuning procedure adheres strictly to the original protocols.

Our configuration. We consider the full spectrum of capabilities in our sampling, excluding OCR, which is reserved to validate capability scalability in Supp. D. While SSR supports diverse judge sources, our main experiments utilize specialized models fine-tuned on 15% GPT-4o annotations to infer on the entire pool. We further explore and analyze alternative deployment strategies across different resource settings in Sec. 4.3.

Baselines. We compare SSR with 10 methods spanning random sampling, perplexity (PPL-mid, PPL-si), the score- and embedding-based LLM selector Deita (Liu et al., 2024c)², VL alignment metrics CLIP (Radford et al., 2021) and E5-V (Jiang et al., 2024), and MLLM SOTAs COINCIDE (Lee et al., 2024), ICONS (Wu et al., 2024), PreSel (Safaei et al., 2025), and PRISM (Bi et al., 2025), all under

¹<https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data>

²Deita’s $O(N^2)$ similarity-threshold diversity filter does not scale to 2.6M instances; we run its score-only variant Deita_{top-k}, ranking by its scores (Supp. B.2).

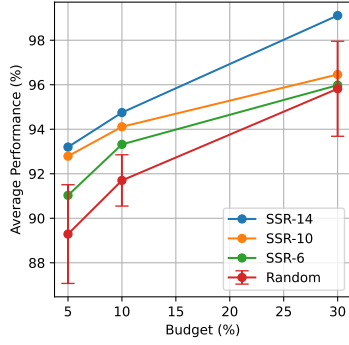


Figure 3: **Richness governs informativeness.** Randomly cutting the capabilities set from full to 10 to 6 monotonically degrades performance, yet they still significantly outperform random sampling. Dimensional richness drives the gain, demonstrating robustness to composition.

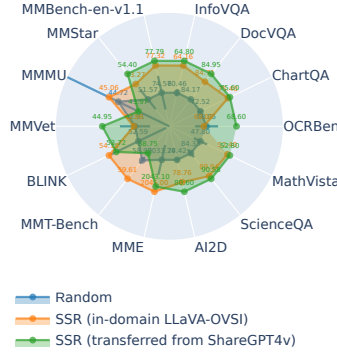


Figure 4: **Training-free cross-domain SSR model transfer.** SSR models finetuned on 10% of ShareGPT4V (12 sources, Vicuna-CLIP) directly sample from the LLaVA-OVSI pool (91 sources, Qwen2-SigLIP), identifying informative data under unseen distributions and target model.

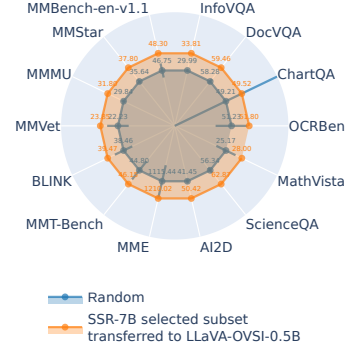


Figure 5: **Cross-model selected subset transfer.** The 5% subset selected by 7B distilled SSR models remains superior when used to train a 0.5B LLaVA-OVSI variant, confirming the explicit profile captures intrinsic data value robust to target-model scale and architecture.

their official configurations (Supp. B.2).

Evaluation benchmarks. We evaluate on 14 multi-modal benchmarks with VLMEvalKit (Duan et al., 2024); the full list and protocol are in Supp. B.1.

4.2 Selection at Million-Scale

Million-scale scalability and benchmark-wide generality. As shown in Tab. 1, among all evaluated methods SSR is the only one that scales to the 2.6M LLaVA-OVSI pool *and* delivers generality across all benchmarks. It recovers 99.1% of full-data performance at 30% and remains robust from cold-start (5%) to warm-start (30%) regimes, regardless of the specific evaluation focus. The $> \text{Rand}$ column makes the generality point sharp: at the 10% and 30% budgets SSR beats the best competitor, i.e., random sampling on **10/10** benchmarks, while every competitor wins on at most 6 of the 10, validating that latents and narrow criteria fail to cover the diverse axes of MLLM evaluation.

Competitor Analysis. LLaVA-OV mid-stage model lacks instruction tuning, rendering its perplexity scores uninformative; the fully fine-tuned SFT checkpoint (LLaVA-OVSI) helps marginally but defeats the purpose of selection, which is to reduce training cost before full fine-tuning. Deita’s textual scorers show that narrow, text-centric metrics such as *quality* and *complexity* do not transfer to the multi-modal domain. VL embedding-based methods (CLIP, E5-V), while capable of encoding both modalities, optimize for contrastive image-text alignment that is misaligned with the generative instruction-following objective of MLLMs. Finally, methods validated on the thousand-scale LLaVA pool all prove brittle to the scale and dis-

tribution shifts of a $4\times$ -larger, heterogeneous pool, confirming that model-anchored cluster or influence signals do not scale and generalize. For example, recent PRISM’s intrinsic-feature scores are computed against a single corpus-level reference and thresholded per sample, so on a pool spanning 90+ sources the ranking reflects global distribution shift more than local redundancy.

4.3 Richness Drives Generality Robustly

Generality stems from richness, not from a particular criterion set or annotator. Why does SSR scale successfully to million-scale and generalize robustly across benchmarks? We attribute it to the *richness* of the capability–style profile, then verify this attribution against two practical perturbations: changes in the criterion set, and changes in the annotator that produced the profile.

Are both score- and style-richness necessary? The SSR profile combines rich capability scores with rich interaction styles. Tab. 2 ablates the two: Score-rich (ScrR) and Style-rich (StyR) each lift performance over random sampling, but only their combination delivers benchmark-wide gains: scores select high-value knowledge while styles ensure that knowledge is delivered through diverse forms. As an ablation, the SP(oor) uses a single *quality* metric; its narrow and inconsistent benchmark coverage confirms that a one-dim criterion cannot cover the diverse axes of MLLM evaluation. Rich scores and rich styles are jointly necessary.

Does richness require an optimal criterion composition? If SSR depended on finding an NP-hard-optimal subset of capabilities, the method would be brittle. Fig. 3 rules this out: randomly cutting the criteria set from full to 10 to 6 capabilities monoton-

ically degrades performance, yet every random subset still significantly outperforms random sampling. What matters is the *richness* of data dimensions, not a singular optimal composition.

Does richness require a high-end annotator? Since the SSR profile is read explicitly from each sample, the framework should decouple from the annotator that produced it. We test two configurations sharing the same selection method and identical query prompts: SSR_{GPT-4o} fine-tunes proxy models on 15% GPT-4o annotations to infer over the full pool, while SSR_{source} in Tab. 2 directly leverages the instruction-following capability of {source} (Qwen2-VL-7B or LLaVA-OVSI-7B). Results show that specialized proxies remain optimal; yet direct querying with open-weight annotators already consistently outperforms the best competing baselines. SSR is thus robust to judge sources, with specialized proxies for peak quality and open-weight prompting for API-cost-free use.

4.4 Cross-Datapool and -Model Transfer

Training-free cross-domain SSR model transfer. SSR distilled scorer and styler models fine-tuned on one architecture with one data pool transfer zero-shot to another, with no retraining and no gradient update. To simulate practical curation under evolving data distributions, we finetune scorer and styler (Vicuna + CLIP) on a 10% random subset of ShareGPT4V (Chen et al., 2024b) containing only 12 sources, then apply them *directly* to infer and sample from the 4 $\times$ -larger LLaVA-OVSI dataset (Li et al., 2025) spanning 91 distinct sources. Fig. 4 shows SSR still identifies high-value data despite the unseen knowledge distribution and target model, because the SSR profile describes sample-intrinsic semantics that remain objective and meaningful when the settings shift.

Cross-model selected subset transfer. High-quality data should be universally beneficial rather than coupled to a particular architecture or training setup. To test this, we feed the subset selected by our 7B distilled SSR models into training a 14 $\times$ -smaller model. As shown in Fig. 5, SSR’s superiority persists, confirming that the explicit profile captures intrinsic data value that is robust to the target model’s scale and architecture.

Zero-shot transfer to a third architecture family at 7M scale. Stressing both transfers at once, we apply the SSR models distilled in Sec. 4.2 *zero-shot* to the unseen 7M Cambrian-7M pool and train Cambrian-1-8B (Tong et al., 2024) on the se-

Cambrian-7M	5%	10%	30%
Random / FULL	90.54	93.57	97.85
SSR (Ours) / FULL	92.40	94.35	98.19

Table 3: **Zero-shot transfer to Cambrian-1-8B on the Cambrian-7M pool.** LLaVA-OVSI-distilled SSR models profile the unseen pool without adaptation; /FULL (%) on the benchmark suite of Tab. 1.

lected subsets under its official recipe (Supp. B.3). Cambrian-1 is fully open across all training stages, permitting a controlled full-data reference, and is disjoint from both stacks above on the vision side (a four-encoder ensemble) and the language side (Llama-3-8B). Tab. 3 shows that SSR outperforms random sampling at every budget (+1.86, +0.78 and +0.34 points at 5%, 10% and 30%), with the largest gain in the cold-start regime where selection matters most, and recovers 98.2% of full-data performance with 30% of the pool. The transfer is demanding along three axes at once, with an unseen pool 2.7 $\times$ larger than the distillation pool, a new source mixture, and a new architecture family, yet needs no re-annotation, fine-tuning, or change to the taxonomy: nothing in the profile depends on the target backbone, so it carries over unchanged.

Prior MLLM selectors are typically validated on a single LLaVA-1.5-era stack with LoRA fine-tuning. SSR’s models and taxonomy transfer unchanged across architecture families, data pools of up to 7M samples, and standard full-parameter SFT: distill a selector once, and it can be reused directly as backbones and pools evolve, precisely the setting where curation decisions are made.

5 Conclusions

We introduced SSR, shifting multi-modal data selection from narrow model-centric signals to rich, sample-intrinsic criteria: each sample is profiled by an interaction style (*how*) and a capability score vector (*what*), and selection is cast as submodular maximization with theoretical coverage and EF1 guarantees. As the first principled multi-modal SFT selector at million-sample scale, SSR delivers benchmark-wide gains across three MLLM architecture families, data pools of up to 7M samples, 6 budgets, and 14 benchmarks, and transfers training-free across data pools and architectures. Its gains stem from the richness of our framework, which is robust to exact composition or annotator, positioning data-centric profiling as a practical lever as MLLM pools outgrow today’s curation pipelines.

Limitations

SSR is validated across three architecture families (LLaVA-1.5, LLaVA-OneVision, Cambrian-1), three data pools of up to 7M samples, 6 budgets, 14 benchmarks, and three complementary judge sources, with formal $(1 - 1/e)$ and EF1 guarantees on the coverage surrogate; we note the boundaries of this evidence. Our target backbones span 0.5B–8B parameters, the range in which controlled full-data references remain affordable and in which prior MLLM selection work is validated. Since the profile carries no model-intrinsic signal, we expect its gains to carry over to larger backbones, but a controlled study at the 72B scale is left to future work. Our profiles depend on an MLLM judge. Gains are consistent across proprietary, open-weight, and distilled judges, and because every score is explicit and human-readable, judge bias is auditable and correctable rather than hidden in a latent space; still, the absolute quality of a selected subset on rare or highly specialized capabilities is bounded by the judge, and we did not sweep judge choices under matched annotation budgets. Likewise, our capability–style taxonomy is broad and open to extension without retraining the selector, but we do not claim it is uniquely optimal, and characterizing its returns on emerging domains (e.g., agentic tool use, long-horizon dialogue) and modalities (e.g., video, audio) is a promising direction we leave to future exploration.

Ethics Statement

This work studies data selection for multi-modal instruction tuning and introduces no new human-subject data collection. All datasets (ShareGPT4V, LLaVA-OneVision, Cambrian) are publicly released by their authors and used here in accordance with their licenses for non-commercial academic research; we did not attempt to re-identify depicted individuals or augment the pools with private data. Because SSR *selects from* pre-existing pools rather than generating new content, any social, demographic, or linguistic biases in the source pool will persist in, and may be reweighted by, the curated subset. The capability–style coverage objective broadens representational diversity along instruction-relevant axes but does not target demographic fairness, and we make no fairness claim along protected-attribute dimensions; downstream users should audit content-level risks separately. We query GPT-4o for distillation labels under Ope-

nAI’s research terms; its outputs are used only as supervisory signals for our distilled SSR models and are not redistributed. SSR is intended as a research tool to make MLLM instruction tuning more efficient and transparent; the interpretability of its capability and style is designed to make misuse easier, not harder, to detect.

Acknowledgements

This work was partly supported by National Natural Science Foundation of China No. 62525103 and Beijing Natural Science Foundation No. L223023.

References

- Georgios Amanatidis, Georgios Birmpas, Philip Lazos, Stefano Leonardi, and Rebecca Reiffenhäuser. 2023. Round-robin beyond additive agents: Existence and fairness of approximate equilibria. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 67–87.
- Jinhe Bi, Zengjie Jin, Yifan Wang, Danqi Yan, Wenke Huang, Xiaowen Ma, Sikuan Yan, Artur Hecker, Mang Ye, Xun Xiao, et al. 2025. PRISM: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. 2016. The unreasonable fairness of maximum Nash welfare. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 305–322.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Sriniwasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024a. AlpaGasus: Training A Better Alpaca with Fewer Data. In *The Twelfth International Conference on Learning Representations*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. ShareGPT4V: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024c. [Are we on the right way for evaluating large vision-language models?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

- Ruibao Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024d. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4156–4172, Bangkok, Thailand. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Harshita Diddee and Daphne Ippolito. 2025. Chasing random: Instruction selection strategies fail to generalize. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1943–1957.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, et al. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2024. What’s in my big data? In *The Twelfth International Conference on Learning Representations*.
- Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Schwarz, Ryutaro Tanno, and Olivier J. Hénaff. 2024. [Bad students make great teachers: Active learning accelerates large-scale visual understanding](#). In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII*, page 264–280, Berlin, Heidelberg. Springer-Verlag.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2025. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [BLINK: Multimodal large language models can see but not perceive](#). In *European Conference on Computer Vision*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, and 15 others. 2023. DataComp: In search of the next generation of multimodal datasets. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. 2025. Metadata conditioning accelerates language model pre-training. In *Forty-Second International Conference on Machine Learning*.
- Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. 2024. Scaling Laws for Data Filtering – Data Curation cannot be Compute Agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22702–22711.
- Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. [E5-V: Universal Embeddings with Multimodal Large Language Models](#). *Preprint*, arXiv:2407.12580.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H erve J egou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.
- Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5060–5080.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.
- Minjae Lee, Minhyuk Seo, Tingyu Qu, Tinne Tuytelaars, and Jonghyun Choi. 2026. [OASIS: Online sample selection for continual instruction tuning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3491–3515, San Diego, California, United States. Association for Computational Linguistics.

- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [LLaVA-OneVision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024a. DataComp-LM: In search of the next generation of training sets for language models. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635.
- Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520.
- Richard J. Lipton, Evangelos Markakis, Elchanan Mossel, and Amin Saberi. 2004. On approximately fair allocations of indivisible goods. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 125–131.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024b. SelectIT: Selective instruction tuning for LLMs via uncertainty-aware self-reflection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024c. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. [MMBench: Is your multi-modal model an all-around player?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024e. OCR-Bench: On the hidden mystery of OCR in large multi-modal models. *Science China Information Sciences*, 67(12):220102.
- Ziche Liu, Rui Ke, Yajiao Liu, Feng Jiang, and Haizhou Li. 2025. [Take the essence and discard the dross: A rethinking on data selection for fine-tuning large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6595–6611, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024f. [Less is More: High-value Data Selection for Visual Instruction Tuning](#). Preprint, arXiv:2403.09559.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024a. #InsTag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. [MathVista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Taffjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Mengyao Lyu, Tianxiang Hao, Xinhao Xu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Learn from the learnt: Source-free active domain adaptation via contrastive sampling and visual persistence. In *Computer Vision – ECCV 2024*, pages 228–246, Cham. Springer Nature Switzerland.
- Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, and Guiguang Ding. 2023. Box-level active detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23766–23775.
- Pratyush Maini, Sachin Goyal, Zachary C. Lipton, J. Zico Kolter, and Aditi Raghunathan. 2024. [T-MARS: Improving visual representations by circumventing text feature learning](#). In *The Twelfth International Conference on Learning Representations*.

- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. 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.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *WACV*.
- Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. 2022. [Towards robust and reproducible active learning using neural networks](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 223–232.
- George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294.
- Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. 2025. Improving data efficiency via curating LLM-driven rating systems. In *The Thirteenth International Conference on Learning Representations*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. [The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yulei Qin, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li, and Xing Sun. 2024. Unleashing the Power of Data Tsunami: A Comprehensive Survey on Data Assessment and Selection for Instruction Tuning of Language Models. *Transactions on Machine Learning Research*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. 2025. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14247–14256.
- Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.
- Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. QuRating: Selecting High-Quality Data for Training Language Models. In *International Conference on Machine Learning (ICML)*.
- Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2024. ICONS: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*.
- Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *Forty-First International Conference on Machine Learning*.
- Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. 2025. [Re-thinking data selection at scale: Random selection is almost all you need](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2698–2711.
- Qian Yang, Shivam Chandhok, Oscar Manas, Kanishk Jain, Aishwarya Agrawal, and Leonid Sigal. 2026. Learning what matters: Prioritized concept learning

via relative error-driven sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, and 3 others. 2024. [MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI](#). In *Forty-first International Conference on Machine Learning*, pages 57116–57198.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. [MM-Vet: Evaluating large multimodal models for integrated capabilities](#). In *Forty-first International Conference on Machine Learning*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36.

A Proofs

This section supplies the formal statements and proofs underlying the guarantees claimed in Sec. 3.3.

A.1 Lifted ground set and partition matroid

To analyze the stratified round-robin algorithm of Sec. 3.3, we lift the pool to the instance–cell pairs

$$V\tilde{V} = \{(i, c, \sigma) : i \in G_{c\sigma}\},$$

and define the lifted coverage objective

$$F\tilde{F}(S\tilde{S}) = \sum_{(c,\sigma)} w_{c\sigma} \min\left\{\tau_{c\sigma}, \sum_{(i,c,\sigma) \in S\tilde{S}} q_{c\sigma}(i)\right\}. \quad (4)$$

Let $V\tilde{V}_{c\sigma}$ denote the block of pairs in cell (c, σ) . The per-cell budgets $\{k_{c\sigma}\}$ induce the partition matroid

$$\mathcal{I}\tilde{\mathcal{I}} = \{S\tilde{S} \subseteq V\tilde{V} : |S\tilde{S} \cap V\tilde{V}_{c\sigma}| \leq k_{c\sigma} \forall (c, \sigma)\}, \quad (5)$$

which is clean regardless of overlap among the original $G_{c\sigma}$. Cells with $G_{c\sigma} = \emptyset$ contribute no pairs and are dropped from the grid, so $\tau_{c\sigma} > 0$ for every remaining cell.

We write $S\tilde{S}_{\text{RR}} \subseteq V\tilde{V}$ for the *idealized* round-robin assignment in which every cell independently claims the top- $k_{c\sigma}$ pairs of its own block $V\tilde{V}_{c\sigma}$ in descending order of $q_{c\sigma}$ (all of them if $|V\tilde{V}_{c\sigma}| < k_{c\sigma}$), under the same fixed tie-breaking rule as the selector; its projection onto the pool is $\hat{D}_{\text{ideal}} = \{i : \exists (c, \sigma), (i, c, \sigma) \in S\tilde{S}_{\text{RR}}\}$. The implemented selector of Sec. 3.3 differs in one respect: a sample selected by one cell is removed from every other queue, so each sample is claimed exactly once. The next lemma transfers guarantees proven for $S\tilde{S}_{\text{RR}}$ to the implemented output $\hat{D}$.

Lemma 1 (Containment). *Let $\hat{D}$ be the pool selected by the implemented stratified round-robin. Then $\hat{D} \supseteq \hat{D}_{\text{ideal}}$, and consequently*

$$F(\hat{D}) \geq F(\hat{D}_{\text{ideal}}) \geq F\tilde{F}(S\tilde{S}_{\text{RR}}).$$

Proof. Fix $(i, c, \sigma) \in S\tilde{S}_{\text{RR}}$, i.e., i ranks within the top $k_{c\sigma}$ of $G_{c\sigma}$. If i is selected by any cell during the implemented run, then $i \in \hat{D}$. Otherwise i stays in cell (c, σ) 's queue throughout; removals by other cells only promote its rank, which starts at $\leq k_{c\sigma}$, and the cell keeps drawing from its non-empty queue until it holds $k_{c\sigma}$ samples, so it would claim i within its first $k_{c\sigma}$ picks: a

contradiction. Hence $\hat{D} \supseteq \hat{D}_{\text{ideal}}$. The first inequality then follows from the monotonicity of F (Prop. 1); the second holds because, cell by cell, $\sum_{i \in \hat{D}_{\text{ideal}}} q_{c\sigma}(i) \geq \sum_{(i,c,\sigma) \in S\tilde{S}_{\text{RR}}} q_{c\sigma}(i)$ with $q \geq 0$ and $x \mapsto \min\{\tau_{c\sigma}, x\}$ non-decreasing.

A.2 Submodularity of F (Prop. 1)

Proof. Each cell term $F_{c\sigma}(S) = w_{c\sigma} \min\{\tau_{c\sigma}, \sum_{i \in S} q_{c\sigma}(i)\}$ is the composition $F_{c\sigma} = \varphi \circ m$ of the non-decreasing concave function $\varphi(x) = w_{c\sigma} \min\{\tau_{c\sigma}, x\}$ with the modular (additive) sum $m(S) = \sum_{i \in S} q_{c\sigma}(i)$; such a composition is monotone submodular (Lin and Bilmes, 2011). Explicitly, for $S \subseteq T \subseteq D$ and $i \notin T$, we have $m(S) \leq m(T)$ since $q_{c\sigma} \geq 0$, and concavity of φ gives

$$\begin{aligned} \varphi(m(S) + q_{c\sigma}(i)) - \varphi(m(S)) \\ \geq \varphi(m(T) + q_{c\sigma}(i)) - \varphi(m(T)), \end{aligned}$$

which is the submodularity inequality for $F_{c\sigma}$; monotonicity follows since φ is non-decreasing and $q_{c\sigma}(i) \geq 0$. Non-negative linear combinations of monotone submodular functions remain monotone submodular. Normalization holds since $F(\emptyset) = \sum_{(c,\sigma)} w_{c\sigma} \min\{\tau_{c\sigma}, 0\} = 0$ with $\tau_{c\sigma} > 0$.

A.3 $(1 - 1/e)$ coverage approximation

Theorem 1 (Round-robin is greedy on a partition matroid; $(1 - 1/e)$). *Let $S\tilde{S}_{\text{RR}}$ be the idealized round-robin assignment of Supp. A.1 and let $S\tilde{S}^* \in \arg \max_{S\tilde{S} \in \mathcal{I}\tilde{\mathcal{I}}} F\tilde{F}(S\tilde{S})$. Then*

$$F\tilde{F}(S\tilde{S}_{\text{RR}}) \geq (1 - 1/e) F\tilde{F}(S\tilde{S}^*),$$

and hence, by Lem. 1, the pool subset $\hat{D}$ selected by the implemented stratified round-robin satisfies $F(\hat{D}) \geq (1 - 1/e) F\tilde{F}(S\tilde{S}^*)$.

Proof. $F\tilde{F}$ has the same saturated-coverage form as F , so the argument of Prop. 1 (composition of a non-decreasing concave function with a modular sum, followed by a non-negative linear combination) applies verbatim on $2^{V\tilde{V}}$; hence $F\tilde{F}$ is monotone submodular. The key structural property we exploit is that $F\tilde{F}$ decomposes additively across the partition blocks, $F\tilde{F}(S\tilde{S}) = \sum_{(c,\sigma)} F\tilde{F}_{c\sigma}(S\tilde{S} \cap V\tilde{V}_{c\sigma})$ with $F\tilde{F}_{c\sigma}$ the (c, σ) summand of Eq. 4, each depending only on pairs inside its own block. The partition-matroid constraint $|S\tilde{S} \cap V\tilde{V}_{c\sigma}| \leq k_{c\sigma}$ decomposes block-wise as well, so maximizing $F\tilde{F}$ over $\mathcal{I}\tilde{\mathcal{I}}$ splits into $|\hat{C}||\hat{\Sigma}|$ independent cardinality-constrained monotone submodular subproblems, one per block. Within each

Method	Vision-Flan	LLaVA-1.5	LLaVA-OVSI
COINCIDE (Lee et al., 2024)	$K=5,000, \tau=0.1$	$K=10,000, \tau=0.1$	$K=20,000, \tau=0.1$
ICONS (Wu et al., 2024)	warm-up 5%, $p=0.2$	warm-up 5%, $p=0.2$	warm-up 5%, $p=0.2$
PreSel (Safaei et al., 2025)	$k=5$	$k=5$	$k=5$
PRISM (Bi et al., 2025)	Layer 1	Layer 1	Layer 1

Table 4: **Baseline configurations.** Settings reported in the original papers (Vision-Flan, LLaVA-1.5) and those used on LLaVA-OVSI; symbols follow each method’s original notation.

block, the idealized assignment selects pairs in descending order of $q_{c\sigma}$; since the marginal gain of pair (i, c, σ) when its block has accumulated mass x is $w_{c\sigma} \min\{q_{c\sigma}(i), \max\{\tau_{c\sigma} - x, 0\}\}$, non-decreasing in $q_{c\sigma}(i)$ for every $\tau_{c\sigma} > 0$ and $w_{c\sigma} > 0$, this order is a valid trajectory of the discrete greedy algorithm for the cardinality-constrained subproblem (ties, including the vanishing gains once a block saturates, broken by a fixed rule; the cyclic order across cells is irrelevant for the idealized assignment since its blocks do not interact). The Nemhauser–Wolsey–Fisher $(1 - 1/e)$ guarantee for cardinality-constrained monotone submodular maximization (Nemhauser et al., 1978) therefore applies in each block, and summation preserves the bound globally. The consequence for $\hat{D}$ is immediate from Lem. 1. Since the argument only uses that the marginal gain is non-decreasing in $q_{c\sigma}(i)$, the bound holds verbatim for every $\alpha \in (0, 1]$ and $w_{c\sigma} > 0$. The benchmark $\tilde{SS}^*$ is the per-cell-budgeted optimum, not the global $b\%$ -cardinality optimum of F ; the per-cell budgets are the coverage mechanism whose fairness Thm. 2 certifies.

A.4 Envy-freeness up to one item

Theorem 2 (Envy-free up to one item across cells). *Treat each cell (c, σ) as an agent with additive valuation $v_{c\sigma}(S) = \sum_{i \in S} q_{c\sigma}(i)$. Let $A_{c\sigma} \subseteq \hat{D}$ be the set of instances claimed by cell (c, σ) during stratified round-robin. Then the allocation $\{A_{c\sigma}\}$ is envy-free up to one item: for every pair of cells (c, σ) and (c', σ') ,*

$$v_{c\sigma}(A_{c\sigma}) \geq v_{c\sigma}(A_{c'\sigma'}) - \max_{i \in A_{c'\sigma'}} v_{c\sigma}(\{i\}).$$

Proof. Each cell-agent’s valuation $v_{c\sigma}(S) = \sum_{i \in S} q_{c\sigma}(i)$ is additive with $q_{c\sigma}(i) \geq 0$. The implemented selector is the classical round-robin picking protocol over a common pool: agents pick once per round in a fixed cyclic order, each claiming the remaining sample it values most (samples outside $G_{c\sigma}$ have value exactly zero to agent (c, σ) , and a claimed sample is removed from every queue), with

equal budgets $k_{c\sigma}$ keeping the sequence balanced. For additive valuations this protocol yields an envy-free-up-to-one-item allocation (the EF1 notion is due to Lipton, Markakis, Mossel and Saberi (Lipton et al., 2004)); the round-robin guarantee is standard, see, e.g., Caragiannis et al. (Caragiannis et al., 2016)); an approximate EF1 analogue holds for the saturated valuation $\min\{\tau_{c\sigma}, v_{c\sigma}(\cdot)\}$ (Amanatidis et al., 2023). For self-containedness, fix agents a and b : if a precedes b in the cyclic order, a ’s round- t pick was made while b ’s round- t pick was still available, so $v_a(A_a) \geq v_a(A_b)$; if b precedes a , dropping b ’s first pick and pairing a ’s round- t pick with b ’s round- $(t+1)$ pick (still available at a ’s turn in round t) gives the stated inequality. An agent whose queue empties early values all remaining samples at zero, so later picks add no envy.

B Additional Experimental Details

B.1 Evaluation Benchmarks

Under the VLMEvalKit (Duan et al., 2024), we comprehensively evaluate our method on 14 multi-modal benchmarks, including MME (Fu et al., 2025), MMBench_{en-v1.1} (Liu et al., 2024d), MMStar (Chen et al., 2024c), MMMU (Yue et al., 2024), MMVet (Yu et al., 2024), BLINK (Fu et al., 2024), MMT-Bench (Ying et al., 2024), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), MathVista_{MINI} (Lu et al., 2024b). For the experiment in Supp. D that scales up in the OCR capability, we additionally evaluate SSR on OCRBench (Liu et al., 2024e), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021) and InfoVQA (Mathew et al., 2022). Since our setup focuses on the single-image SFT phase, the model does not possess the multi-image understanding ability. Thus, for MMMU and BLINK, we report results on the single-image QA split.

B.2 Baselines

We compare SSR with 10 methods, including a) *random sampling*: the strong diversity-prioritized baseline, evaluated based on the average results

from three trials of different random splits; b) *perplexity*, including its two variants before (PPL-mid) and after (PPL-si) the single-image SFT on the entire data pool; c) *Deita* (Liu et al., 2024c), the score- and embedding-based selection method for LLMs, run in its score-only top- k form ($Deita_{top-k}$): Deita controls diversity by thresholding pairwise embedding similarity, whose $O(N^2)$ complexity and dataset-specific threshold do not scale to our 2.6M instances, and any threshold we chose on its behalf would be equally contestable, so we rank by its quality and complexity scores and label the variant explicitly; d) *CLIP similarity* (Radford et al., 2021) (ViT-L) that evaluates the image-text alignment; e) *E5-V similarity* (Jiang et al., 2024), the SOTA MLLM-based universal embedding model built on LLaMA-3-8B (Dubey et al., 2024); f) COINCIDE (Lee et al., 2024), ICONS (Wu et al., 2024), and PreSel (Safaei et al., 2025), the SOTA selection strategies for MLLMs; and g) *PRISM* (Bi et al., 2025), the most recent selector, which scores each sample by its correlation with the corpus-mean-centered intrinsic features of the target MLLM and prunes by a global threshold.

All baseline configurations follow settings endorsed by their original authors; Tab. 4 lists the complete choices. For COINCIDE, the two hyperparameters are the number of clusters K and the temperature τ of its cluster-level sampling distribution, with the official $\tau=0.1$. For ICONS, the warm-up ratio fixes the fraction of the pool used to train the warm-up model underlying its influence scores, and p sets the proportion of top samples per task: a sample passes the consensus vote for a task when its influence reaches the task-specific threshold τ_k , the $((1-p) \times 100)$ -th percentile of that task’s influence distribution. We adopt its best-performing 5% warm-up ratio (small warm-up sets consistently outperform the full pool in their ablation) and the main-experiment $p=0.2$, i.e., 80th-percentile thresholds. For PreSel, k is the number of nearest neighbors in feature space for computing the NC score, reported robust across $k \in \{5, 10, 20\}$ and set to 5 in all its experiments. For PRISM, intrinsic visual features are extracted at the shallow Layer 1, which its authors find most effective for redundancy detection. The competitors were designed for thousand-scale pools; since the settings above are either scale-free or reported robust, we retained each method’s official configuration. COINCIDE alone provides scale-adaptation guidance (K grows from 5,000 on the 186K Vision-Flan pool to 10,000 on the 665K

LLaVA pool, a $\approx 3.5\times$ increase), which we extrapolated to $K=20,000$ on the 2.6M pool. We emphasize that the competitors’ difficulty at million scale stems from their model-centric design and narrow criteria (Sec. 4.2) rather than from hyperparameter sensitivity: SSR itself exposes no tunable selection hyperparameter (Sec. 3.3), so no tuning asymmetry favors our method.

B.3 Cambrian-1 Transfer Protocol

For the third-architecture transfer in Sec. 4.4 (Tab. 3), we take the scorer and styler distilled on 15% of LLaVA-OVSI (Sec. 4.2) and apply them *zero-shot* to profile all 7M samples of the Cambrian-7M pool (Tong et al., 2024), with no re-annotation, no fine-tuning, and no change to the taxonomy. The stratified round-robin selector then draws 5%, 10% and 30% subsets, on which Cambrian-1-8B is fine-tuned from its official pre-trained checkpoint with the official recipe³; random sampling and the full-data reference use the identical recipe. Evaluation follows Supp. B.1.

B.4 Profiling Details

This subsection expands the profiling pipeline of Sec. 3.1. Supp. B.4.1 documents how the capability and interaction-style taxonomies were derived; Supp. B.4.2 details the judge prompt that yields the rich per-sample profile consumed by selection.

B.4.1 Capability and Style Taxonomies

Tab. 5 lists the $|C|$ capabilities and $|\Sigma|$ interaction styles we identify from the current open-access multi-modal instruction-following datasets for building general-purpose large models. Our taxonomy is constructed through a three-stage, transferable procedure: (i) *discovery* of an initial set of capabilities from analyzing the LLaVA-OVSI data sources; (ii) *expansion* of this set by prompting GPT-4o with random samples to surface novel patterns; and (iii) two iterations of *refinement* that merge semantically similar items and prune long-tail capabilities. The prompt used in this capability-definition process is shown immediately below in the boxed insert. A parallel methodology yields the taxonomy of interaction styles.

We emphasize that the efficacy of SSR is not rigidly tied to this specific capability composition. As demonstrated in Sec. 4.3, the success of

³https://github.com/cambrian-mlm/cambrian/blob/main/scripts/cambrian/finetune_cambrian_8b.sh

Prompt to Derive the Capabilities of Interest

You are an AI expert tasked with defining the essential capabilities for a next-generation multi-modal large language model. To achieve this, we will follow a structured, three-step process: Discovery, Expansion, and Refinement.

Stage 1: Discovery

First, I want you to act as a research analyst. Your task is to analyze the following data sources to identify a broad set of foundational capabilities and iterative tasks suggested by the data.

Data Sources: [
 COCO Caption,
 Vision FLAN,
 ...
]

Based on your analysis of these sources, generate a list of potential capabilities.^b

Step 2: Expansion

Now, I want you to think beyond this initial list. I will provide you with a list of initial capability candidates and random samples of visual question answer pairs. Your task is to think creatively and identify any new capabilities that are not adequately covered by the existing list but are necessary for answering these visual questions.

Initial List:
{initial capabilities}

Random Samples:
{Random samples drawn from the datapool}

Your goal is to heuristically expand our list. Think creatively and do not worry about overlap at this stage; focus on generating a comprehensive set of potential new capabilities.^c

Step 3: Refinement

Next, I want you to refine our expanded list of capabilities. I will provide you with a list of capabilities. Your task is to identify and merge criteria that are semantically similar or redundant. For each proposed merger, provide a brief justification. The goal is to create a comprehensive yet more concise and semantically coherent list.

Expanded List:
{expanded capabilities}

Present the new, merged list.^d

Next, I want you to analyze the frequency and importance of each capability on the merged list. Your task is to analyze the merged list of capabilities below and identify capabilities that are the most critical for a wide variety of multi-modal questions. Also, identify the potentially long-tailed datapoints, which might be capabilities that are too niche, rarely required, or could be considered a sub-component of other capability we've already defined.

Merged List:
{merged capabilities}

Your goal is to create a final, core set of capabilities for training while discarding less representative ones. Please provide the final list of essential capabilities and a separate list you chose to set aside.^e

^a See Tab. 7 for the full list.

^b The initial list generated by the LLM in this round will be denoted as {initial capabilities}.

^c This step can be iterated multiple times to get a sufficient candidate list. The expanded list generated in this round will be denoted as {expanded capabilities}.

^d The merged list suggested in this round will be denoted as {merged capabilities}.

^e This step can be iterated multiple times to get a refined list.

SSR is predicated on the richness of the capability subspace, where even randomly sampled subsets of capabilities achieve superior performance. Furthermore, our analysis in Supp. D and Fig. 9 shows that the framework seamlessly and effectively scales to incorporate new capabilities. This inherent extensibility ensures that as the data landscape evolves, either through new public datasets or private, specialized sources, our pipeline can be readily adapted to curate data for a broader and more sophisticated range of knowledge. This includes advanced skills such as creative image-based writing, complex chain-of-thought reasoning, and solving competition-level mathematical problems.

B.4.2 Querying Rich Profiles

Prompt template. Given that obtaining large-scale human annotations for both scores and styles is often impractical, prompting multi-modal large language models (MLLMs) serves as a practical and effective alternative. The boxed insert below shows our prompt template for querying rich scores and styles for each sample, where **{Input}** and **{Response}** are questions and answers. Multi-round

user-assistant interactions are paired and concatenated for demonstration.

Annotation protocol. To enhance the stability of pointwise scores rated by the model, we define a scale from 0 to 5 and establish clear benchmarks. During the evaluation process, each data sample is assessed against all predefined capability dimensions, and all interaction styles present in the textual component are recalled. To enhance the reliability and self-consistency of these annotations, we instruct the model to provide a justification for each score it assigns. A critical aspect of our methodology, particularly for multi-modal data, is ensuring a holistic evaluation. Our prompts specifically direct the model to assess the synergy and alignment between the image and text in the context of the task. This is crucial for preventing uni-modal bias, where the model’s judgment might be disproportionately influenced by either the visual or the linguistic content alone.

Annotator choice. As established in Sec. 4.3 of our main paper, the SSR framework is robust to various annotation sources, demonstrating strong performance with both open-weight models and

multi-modal Capabilities	Definitions	Examples
activity recognition	actions or behaviors of humans, animals, or objects	Fig. 15
causal reasoning	cause-and-effect relationships between events or variables to predict outcomes and explain phenomena	Fig. 16
humanities	history, literature, philosophy, art, and culture to understand human experiences and societal developments	Fig. 17
STEM knowledge	science, technology, engineering, and mathematics, chemistry, economics etc	Fig. 18
comparative analysis	compare multiple entities, concepts, or datasets to identify similarities, differences, and relationships	Fig. 19
data understanding	documents, tables, charts, graphics, infographics	Fig. 20
object spatial understanding	the positions, orientations, countings and relationships of objects	Fig. 21
attribute identification	various characteristics and properties of objects, such as identity, color, size, shape, material, emotion, and other distinguishing features	Fig. 22
logical deduction	to analyze information, recognize patterns, draw valid conclusions based on structured principles of logic and make reasoned decisions	Fig. 23
scene understanding	complex environment with objects, their attributes, spatial relationships, and activities, as well as surrounding information and circumstances within the scene	Fig. 24
fine-grained recognition	subtle differences and specific features within similar categories of objects	Fig. 25
language generation	generate coherent and contextually appropriate text in various languages, styles, and formats based on instructions	Fig. 26
in-context learning	follow the demonstrations of the task within a given conversation	Fig. 27
optical character recognition	the conversion between images of printed/handwritten text and machine-readable text	Fig. 28
style	multi-choice, coordinate, yes/no, word/short-phrase, short description, detailed description, comparison, chain-of-thought (step-by-step), specified style	Fig. 15 - 28

Table 5: Criteria we recognize as the foundational pillars for developing vision perception and reasoning capabilities within MLLMs, and the interaction styles we identify from instructional multi-modal data.

Prompt to Query GPT-4o for Rich Scores and Styles

System Prompt:

You are an AI expert rater designed to analyze the Visual Question Answering (VQA) instance in the user query to perform the following tasks step-by-step:

Step 1: Classify the VQA instance into given conversation style.

Step 2: Evaluate the helpfulness of the information provided in the VQA instance with respect to various model capabilities. Specifically, rate how well this information could enhance each capability of a multi-modal large language model through learning from it.

Step 3: Output the results strictly follow the JSON format.

User Prompt:

Instruction

You need to perform the following three steps to rate the User Query and output result in the dictionary format.

Step 1: Classify the instance in interaction style. Determine the task style of the VQA instance and select styles from the list "task_styles" below. Sort the selected styles by frequency of occurrence.

Step 2: Rate each capability from 0-5. For each capability listed and explained in "task_capabilities" below, analyze how effectively the VQA instance could enhance that capability of a Multimodal Large Language Model (MLLM) by learning from it. Rate each capability using the scores from the "score_scale" list below in reference to the guidelines. Please ensure that the scores are well-distributed across the range.

Finally, output the results strictly following the dictionary format defined in Output Format. Do not output any additional tokens outside it.

User Query

Question: {Input}

Answer: {Response}

Task Styling

```
task_styles = [  
    multi-choice,  
    coordinate,  
    ...  
]
```

^a

Task Capabilities

```
task_capabilities = [  
    optical character recognition, # the conversion between images of printed/handwritten text and machine-readable text  
    ...  
]
```

^b

Rating Scale

```
score_scale = [  
    0, # Not Relevant: The VQA instance does not present or relate to the capability in any meaningful way.  
    1, # Minimal: The VQA instance offers very little information relevant to the capability, providing negligible value for enhancement.  
    2, # Fair: The VQA instance contains some relevant information but lacks depth and clarity, contributing minimally to the model's learning in this capability.  
    3, # Good: The VQA instance provides a fair amount of relevant information, which can moderately aid in the model's learning and enhancement of the capability.  
    4, # Significant: The VQA instance offers substantial information that is highly relevant and beneficial, significantly aiding the model's learning and enhancement of the capability.  
    5, # Excellent: The VQA instance is exceptionally rich in relevant information, providing comprehensive and clear insights that would greatly enhance the model's learning and mastery of the capability.  
]
```

Output Format

```
{  
    "style": "<list of string>",  
    "capability2score": "<dict of str:int>",  
    "capability2explanation": "<dict of str:str>",  
}
```

^aSee Tab. 5 for the full list.

^bSee Tab. 5 for the full list and explanation.

the proprietary one. To rigorously evaluate the efficacy of our prompting strategy and the fidelity of the annotations it produces, we focus this analysis on our primary configuration, which leverages a leading proprietary model, i.e., GPT-4o in our case, as the expert annotator. This choice achieves SOTA performance, while the strategy of annotating only a small data subset makes it highly cost-effective. For qualitative illustration, we provide examples for each capability in Fig. 15–28. These examples are organized by capability, showcasing the full 0–5 scoring scale, and are accompanied by detailed rationales for each assigned score.

Case studies. Our scoring mechanism demonstrates a strong sensitivity to task context. Consider, for instance, the 3rd and 5th samples in Fig. 28. Although the visual scenes are similar, the textual queries focus on distinct elements. When the task is to generate an “informative summary” requiring text from a vehicle, the sample’s contribution to the OCR capability is crucial, meriting a score of 3. Conversely, when the task shifts to global scene understanding where road details are paramount, the text becomes irrelevant, and the sample rightly receives an OCR score of 0. Furthermore, our process effectively penalizes samples containing flawed or hallucinatory responses in the original dataset. For example, the hallucination present in the 4th example of Fig. 21 is identified and results in a lower score, thereby preventing the propagation of erroneous information into the supervised fine-tuning (SFT) phase. While model-generated annotations can occasionally include hallucinations (e.g., 5th example in Fig. 21, 5th example in Fig. 28), we find that our finetuning process is remarkably robust to this low level of label noise, as validated by our accuracy analysis and empirical selection results.

Collectively, these cases validate the effectiveness of our task-aware prompting, demonstrating that our valuation criteria successfully balance the intricate interplay between image, query, and task context. Finally, this initial annotation phase, leveraging a small, high-quality set of judgments from GPT-4o, provides the foundation for fine-tuning the specialized SSR models, which are analyzed in detail in Supp. C.

API cost of the 15% GPT-4o annotation phase. For practitioners reproducing the distilled pathway, we report the token accounting of the one-time GPT-4o annotation on 15% of LLaVA-OVSI. The text portion of the prompts totals

428.33M input tokens. Images were submitted with `detail="auto"`; since the server-side routing rule is not disclosed, we bound the image cost between all-low-detail (33.76M tokens) and all-high-detail under the official tile formula (427.40M tokens), giving 462.09M–855.73M total input tokens. The structured JSON outputs total 218.43M tokens. At GPT-4o list prices (\$2.50/M input, \$10/M output), the standard-API cost is $\approx$ \$3.34K–\$4.32K, and roughly half of that ($\approx$ \$1.67K–\$2.16K) with the Batch API. This is a one-time expense: the distilled models profile the remaining 85% locally, adapt zero-shot to new pools with no further API calls (Sec. 4.4, Supp. B.3), and can be bypassed altogether by prompting an open-weight judge (Tab. 2).

C Fidelity of Distilled SSR Models

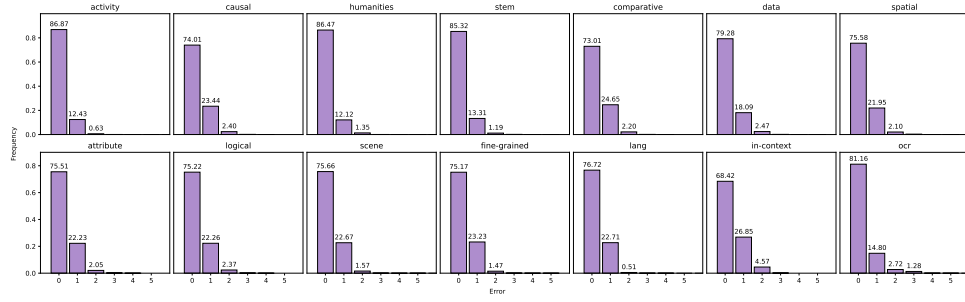
The distilled judge pathway introduced in Sec. 3.4 amortizes proprietary-API cost by fine-tuning small student MLLMs on a 15% judge-labeled subset. This section validates the fidelity of those students against the GPT-4o judge on the remaining 85%, both quantitatively (Supp. C.1) and qualitatively (Supp. C.2).

C.1 Quantitative Validation

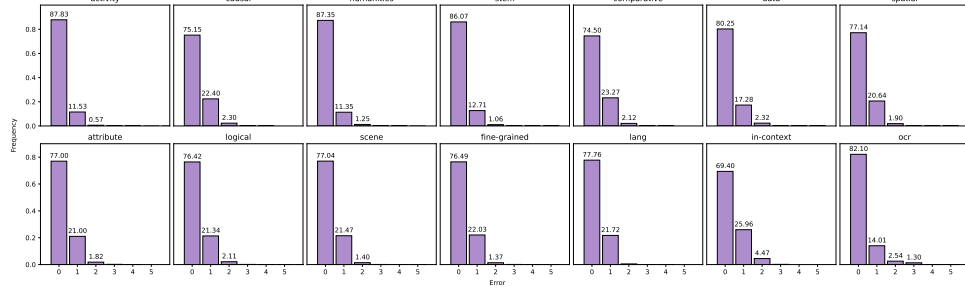
In our primary configuration, the fine-tuning of SSR models uses 15% of the LLaVA-OVSI data, following the original instructional tuning strategy (Li et al., 2025)⁴. To minimize the exploration cost of SSR in practical applications, no hyperparameter search was conducted in fine-tuning. To validate the fidelity of our specialized model, we established a validation set by annotating the remaining 85% of the LLaVA-OVSI data pool using GPT-4o.

Scorer Fidelity. The scorer’s performance, quantified by Mean Absolute Error (MAE), is detailed in Fig. 6(a). Our model demonstrates high alignment with the judge annotator. Averaged across all capabilities considered in the paper, spanning varying levels of granularity and difficulty, our scorer achieves an exact match with GPT-4o on 77.7% of predictions. The agreement rate rises to an impressive 97.8% when allowing for a tolerance of ± 1 . This demonstrates that while pointwise judgments from the judge and our fine-tuned model exhibit inherent stochasticity (Wettig et al., 2024), much like human evaluation, SSR rarely makes unrea-

⁴https://github.com/LLaVA-VL/LLaVA-NeXT/blob/main/scripts/train/finetune_si.sh

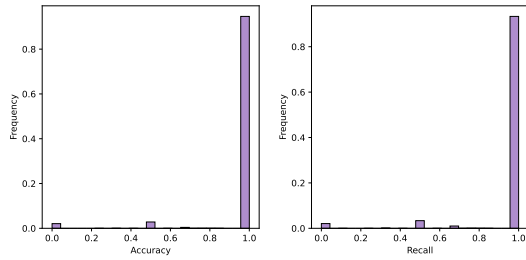


(a) SSR scorers trained with 15% data



(b) SSR scorers trained with 30% data

Figure 6: The mean absolute error of SSR scorer predictions against GPT-4o judgment over capabilities.



(a) Precisions of SSR styler (b) Recalls of SSR styler

Figure 7: The precision and recall of SSR styler predictions against GPT-4o judgment.

sonable decisions. Consequently, it preserves the overall quality stratification essential for effective data selection.

Styler Accuracy. Fig. 7 reports the precision and recall distributions of the styler’s predictions benchmarked against GPT-4o annotations. Compared to fine-grained capability scoring, the classification of interaction styles proves to be a more distinct and less ambiguous task. Consequently, the model achieves exceptional performance, registering an average precision of 96.35% and a recall of 95.80% across the entire data pool.

Annotation Volume Analysis. We further explored the relationship between judge annotation volume and specialized model fine-tuning performance by scaling the fine-tuning set to 30% of the data pool.

The resulting MAE comparison in Fig. 6(b) reveals that performance improvements over the 15% baseline are negligible. This validates the data efficiency of our approach, demonstrating that the scorer achieves effective task alignment with an affordable API cost.

Score Distribution Analysis. We visualize the capability-wise score distributions in Fig. 8. For rare and specialized domains such as humanities and STEM, the model exhibits high precision, particularly in recognizing the presence or absence of a capability (i.e., correctly assigning a score of 0). This sensitivity is crucial for our diversity-oriented round-robin sampling, ensuring that scarce, high-value data points are reliably preserved and not overlooked.

Collectively, these quantitative analyses confirm that our specialized SSR models deliver highly reliable judgments. This fidelity underpins the robust and superior performance of SSR observed in both in-domain inference and direct transfer to new data pools.

C.2 Qualitative Analysis

For each capability, we employed stratified random sampling based on the GPT-4o scoring scale (0–5) to present examples. The resulting visualizations, comprising image-text pairs, GPT-4o scores, identified styles, explanations, and our SSR judgments,

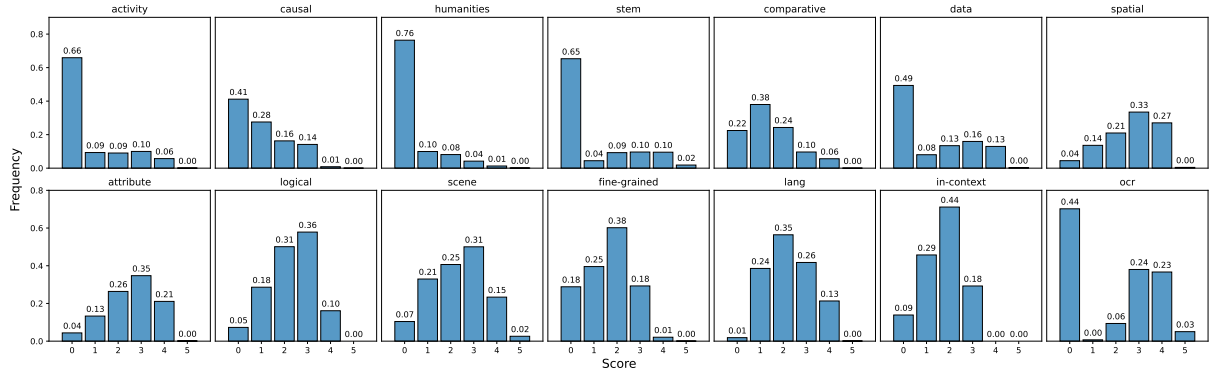


Figure 8: Score distribution of capabilities across the LLaVA-OVSI dataset inferred by SSR.

are presented in Fig. 15–28.

The consistent quality observed across these randomly sampled instances underscores the robustness and fidelity of our SSR judgments. Qualitative inspection further confirms that, for both the proprietary judge and our fine-tuned scorers, discrepancies between adjacent scores often reflect negligible semantic differences that are indistinguishable even to human evaluators. For instance, the distinction between logical reasoning scores of 1 and 2 in the 4th example of Fig. 23, or between comparative analysis scores of 4 and 5 in the 1st example of Fig. 19, is subtle. Thus, while the model may occasionally diverge on the exact integer, it accurately captures the semantic caliber of the data, ensuring robust quality stratification.

D Capability Scaling and Customization

Beyond the data-*quantity* scalability already shown on the 2.6M pool in Sec. 4.2, real-world curation often demands *adding* new capabilities into an existing pool. We use OCR as a held-out case study; a comprehensive sweep across all 14 capabilities would require substantially more compute and is outside the scope of this work, and we do not claim that all capabilities customize identically.

Quantity axis. Along the quantity axis (Fig. 9), SSR delivers consistently superior MLLM performance from extreme cold-start (1%) to resource-abundant (40%, 50%) budgets, extending the million-scale scalability of Sec. 4.2 to a wider budget range.

Capability axis. Along the capability axis, we elicit OCR-specific judgments and let SSR prioritize high-scoring OCR samples: the added samples yield steady improvements on OCR-related benchmarks while maintaining (or further extend-

ing) gains on general benchmarks. SSR’s modular profile therefore supports targeted capability expansion without sacrificing prior coverage.

E Efficiency Comparison

Fig. 10 presents a time-performance analysis, comparing multi-modal data selection methods against the full fine-tuning baseline on LLaVA-OVSI. For each method, the total wall-clock time encompasses both the data selection phase and the subsequent training of the target model on the selected budget. The results demonstrate that SSR significantly outperforms all baselines in efficiency and effectiveness. The gains are particularly striking when compared to full fine-tuning, which requires 62.35 hours on 64 H100 GPUs. Even when executing the entire pipeline from scratch, SSR achieves 93.20%, 94.75%, and 99.11% of the full performance within 30.4, 31.8 and 36.8 hours respectively. This demonstrates that our method can recover nearly all of the performance of the full dataset while reducing the end-to-end training time by over 40%.

A key architectural advantage of explicit data valuation is its efficiency. While each data point of SSR in Fig. 10 accounts for the end-to-end time, the two primary computational costs, model fine-tuning and data valuation, are one-time, amortizable costs. In practice, the fine-tuning of the scorer/styler can be entirely obviated by directly prompting an open-weight MLLM or by directly applying our pre-trained models to a new domain, validated in Sec. 4.3 of the main paper. Data valuation is also a one-time computation, allowing near-instantaneous adaptation to different budget settings through a simple re-ranking of the pre-computed results. This stands in sharp contrast to

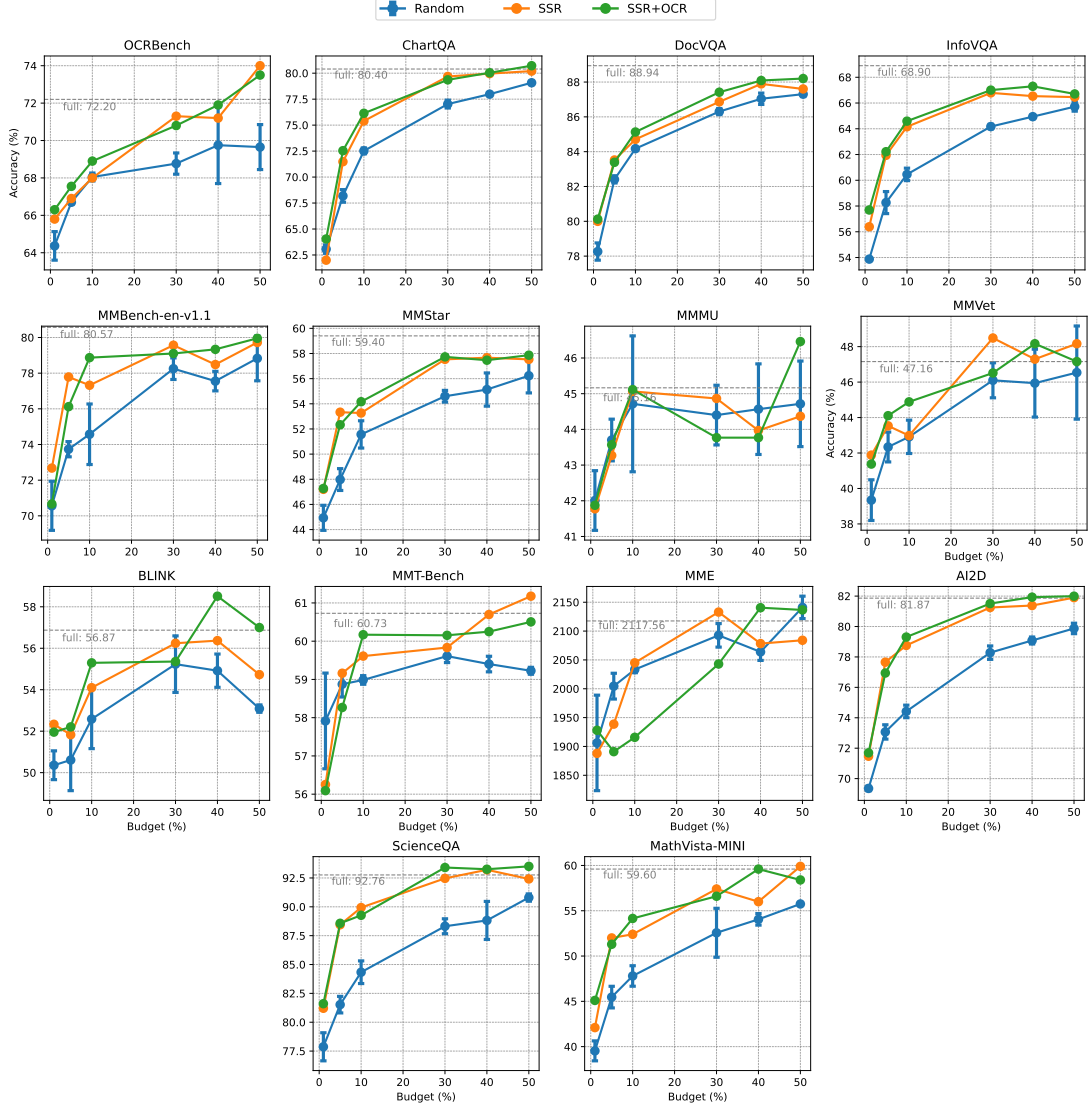


Figure 9: **Scalability along both quantity and capability axes.** SSR consistently surpasses random across six data quantities (1% $\rightarrow$ 5% $\rightarrow$ 10% $\rightarrow$ 30% $\rightarrow$ 40% $\rightarrow$ 50%); adding OCR as a held-out capability (13 $\rightarrow$ 13 + OCR) extends the criteria set without degrading general performance, evidencing modular capability scalability. Legend: Random, SSR without OCR, SSR with OCR; 14 benchmarks.

prior work. COINCIDE (Lee et al., 2024), for example, must re-run its time-consuming clustering and selection process (over an hour on million-scale datasets) for each new selection task. The inefficiency of ICONS (Wu et al., 2024) is even more pronounced, as it necessitates computing gradients on validation data for each target benchmark, followed by a large-scale similarity estimation against the entire candidate pool. This makes its approach not only computationally prohibitive but also fundamentally tied to specific validation sets.

F Validation on Benchmark Test Splits

The comparative results presented in Tab. 1 of the main paper were conducted within a unified framework under identical data budgets across our method and all baselines, ensuring a fair comparison. Inside this batch, the relative ordering between methods is already sufficient to validate the effectiveness of each sampling strategy. To further substantiate the effectiveness of SSR, we additionally evaluate the top competitors on the available test splits that require online assessment, specifically MMBench_{en-v1.1} (Liu et al., 2024d), MMMU (Yue et al., 2024), and MMT-Bench (Ying et al., 2024).

	MMBench _{en-v1.1-test}	MMMU _{test}	MMT-Bench _{all}	>Rand	/FULL
Budget: 5%					
Random	73.45	40.37	59.98	-	96.32%
Deita	<u>73.97</u>	36.30	56.57	1/3	91.40%
COINCIDE	73.09	40.30	57.47	0/3	94.74%
ICONS	69.04	37.30	57.49	0/3	90.63%
SSR	<u>75.84</u>	<u>41.30</u>	<u>60.10</u>	3/3	<u>98.15%</u>
Budget: 10%					
Random	74.55	40.40	60.54	-	97.12%
Deita	<u>75.17</u>	37.00	57.40	1/3	92.92%
COINCIDE	74.44	<u>40.40</u>	58.68	1/3	96.05%
ICONS	71.90	38.80	58.66	1/3	93.68%
SSR	<u>76.05</u>	<u>40.90</u>	<u>60.68</u>	3/3	<u>98.23%</u>
Budget: 30%					
Random	77.33	41.13	59.59	-	98.36%
Deita	76.88	40.00	59.56	0/3	97.24%
COINCIDE	<u>78.18</u>	41.00	<u>59.77</u>	2/3	98.71%
ICONS	72.21	40.40	<u>59.70</u>	1/3	95.67%
SSR	<u>78.13</u>	41.10	<u>59.80</u>	2/3	<u>98.78%</u>
FULL					
LLaVA-OVSI	79.27	41.40	60.70	-	100%

Table 6: Performance comparison on the benchmarks with online test splits conducted across varying budgets of 5%, 10% and 30% of LLaVA-OVSI. We highlight the best result in **boldface** and underline the result if it beats the random baseline. The column >Rand presents the number of benchmarks where the method exceeds random sampling, and /FULL compares the performance of sampled data with that of the FULL dataset.

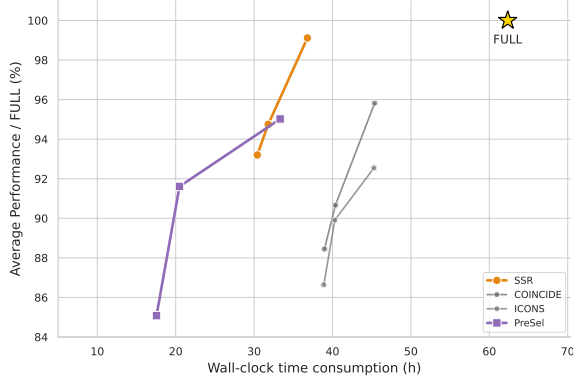


Figure 10: Comparison of selection methods on average performance and total time consumption.

Consistent with the trends from Sec. 4.2, Tab. 6 shows that SSR not only consistently outperforms or matches the competitors across all data budgets, but its advantage is particularly pronounced in the challenging low-data regime. Unlike methods that rely on model-specific features or pre-heated fine-tuning, our approach is grounded in explicit capabilities and interactional styles, which translates into superior efficiency and robustness on the held-out test splits. It is worth noting that the trends on MMMU suggest a performance ceiling for the current data pool. Achieving further breakthroughs

on this benchmark will likely necessitate augmenting the training data with external sources rich in college-level, multidisciplinary knowledge, an extension that the capability-extensible SSR framework readily supports (Supp. D).

G Analysis of Selected Subsets

Having validated SSR’s transferability (Sec. 4.4), scalability (Supp. D), efficiency (Supp. E), and effectiveness (Supp. F), we now look *inside* the selected subsets to understand *why* the selection works. We examine the distribution of the resulting subsets along three complementary axes: capability scores (Supp. G.1), interaction styles (Supp. G.2), and data sources (Supp. G.3).

G.1 Score Distribution

To illustrate the information distribution of the data obtained by our SSR sampler, Fig. 11 presents the score distributions for subsets at 10% and 30% sampling ratios. These statistics are derived from the predictions of the SSR model (trained on 15% annotation data) following our main configuration.

As depicted in Fig. 11(a) and Fig. 11(b), the sample-aggregated capability scores exhibit a distinct bell-shaped curve. This distribution pattern arises from the natural scarcity of samples with maximum scores. After exhausting the top tier, the

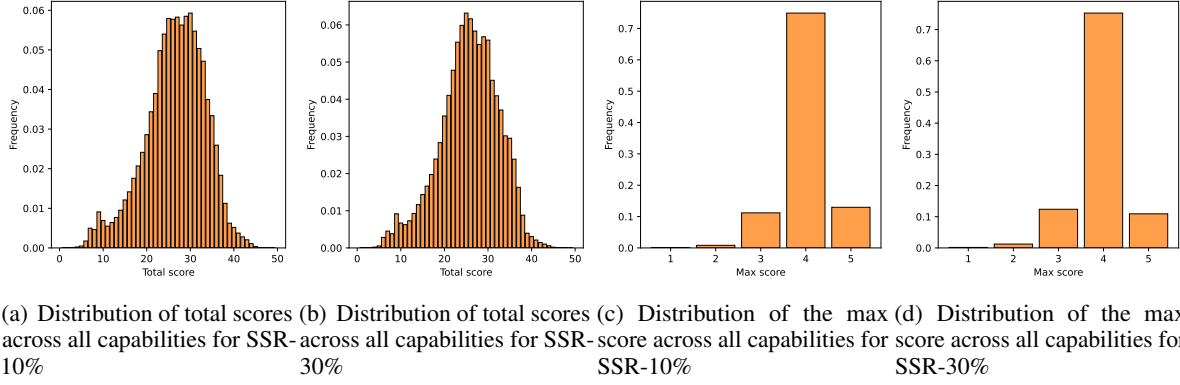


Figure 11: Sample-level score distribution of the selected SSR-10% and SSR-30%.

sampler satisfies the remaining budget with mid-range samples, converging toward a normal distribution centered on the high-quality midrange. Notably, our sampler selects data points even if their aggregate contribution is low, provided they excel in specific dimensions. This behavior is critical because specialized capabilities (e.g., STEM) often have less correlation with other skills, resulting in lower total scores for these “specialist” samples. By coupling an explicit capability subspace with a round-robin selection mechanism, we specifically target these minority dimensions, preventing them from being overshadowed by generalist samples. By enforcing coverage across all dimensions, our selected subsets can address the niche challenges found in complex benchmarks, thereby broadening the model’s operational spectrum. Fig. 11(c) and Fig. 11(d) provide further insight into our sampler’s strategy by illustrating the distribution of sample-wise maximum capability scores. The pronounced concentration of scores around 4, coupled with the deliberate inclusion of lower-scoring samples, underscores SSR’s capacity to strike a critical balance between information efficiency and capability richness.

By integrating these robust specialized capabilities alongside peak-value samples, SSR ensures a rich and efficient capability representation, ultimately enhancing the final model’s overall performance and adaptability.

G.2 Style Distribution

Fig. 12 presents a comparative analysis of interaction style distributions. As illustrated, SSR exhibits a distinct pattern characterized by a balanced representation across diverse styles. This comprehensive coverage is critical for the objective of

the SFT stage, directly contributing to the robustness of the final model. In contrast, other sampling methods display skewed distributions, disproportionately favoring specific styles such as *detailed description* (which inflates token counts) or *yes/no* and *short-phrase* answers (ubiquitous in evaluation benchmarks). Such imbalances can limit the versatility and generalization of the resulting models. Notably, SSR achieves a more equitable distribution, successfully preserving complex styles like *comparison* and *chain-of-thought* that are vital for reasoning tasks. This balance highlights our method’s capacity to support a broad spectrum of downstream applications, establishing SSR as a general-purpose tool for effective dataset curation.

G.3 Data Source Distribution

Given the inherent heterogeneity of multi-modal data sources and the rigorous demands of comprehensive evaluation benchmarks, maintaining diversity during instruction fine-tuning is paramount. Adhering to the original data hierarchy defined in (Li et al., 2025), we present a comparative statistical breakdown of the full data pool versus our 10% sampled subset in Tab. 7, with corresponding visualizations in Fig. 13.

The selected subset reveals a distinct shift toward a more balanced distribution under SSR. As shown in the original LLaVA-OVSI (left), the dataset is dominated by the General category (42%), with COCO Caption alone accounting for 35%. In contrast, the SSR-sampled subset (right) achieves a significantly more equitable distribution. The General category is modulated to 35%, with COCO Caption reduced to 14%, while high-value categories see substantial growth: Math/Reasoning expands from 24% to 29%, and Doc/Chart/Screen increases from

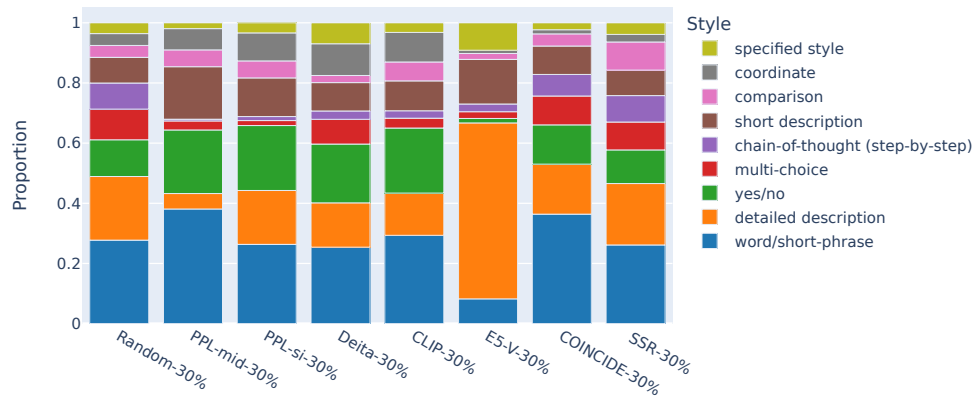
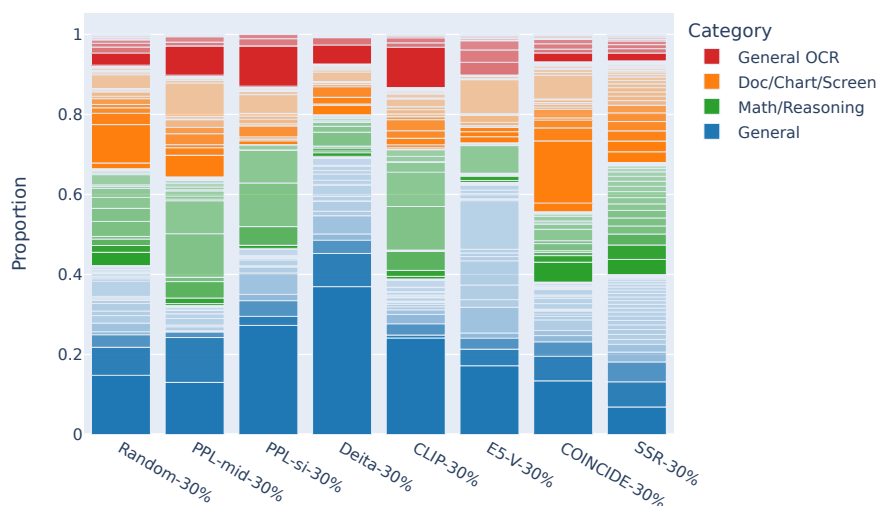
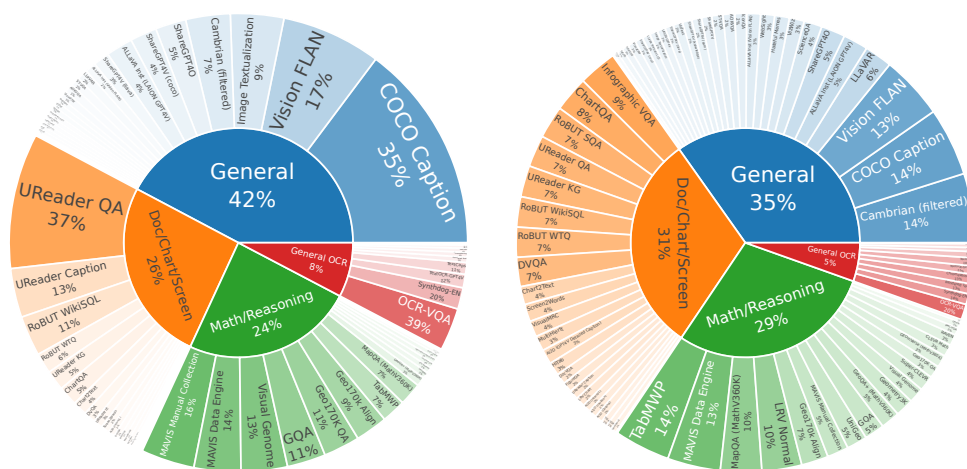


Figure 12: Comparison of data style statistics between our selection and those of competitors.

26% to 31%.

Fig. 14 further contrasts our approach with comparative methods. While SSR maintains balance across diverse sources, baseline methods exhibit clear biases: Deita and E5-V remain heavily concentrated in dominant general data; PPL and CLIP disproportionately favor math/reasoning (especially Visual Genome); and COINCIDE is skewed toward document/chart data. The effective reallocation of the data budget by SSR underscores its ability to achieve a diverse and representative training set. This diversification directly enhances the target model’s robustness on general instruction-following while boosting adaptability for challenging tasks such as mathematical problem-solving and infographic reasoning.

In conclusion, the distribution analysis presented in this section confirms that SSR effectively provides highly informative subsets spanning rich capabilities and diverse instruction styles. This approach significantly advances both the data efficiency and curation of MLLM instruction tuning.



Source	LLaVA-OVSI		SSR-10%		SSR-30%	
	# Samples	Prop.	# Samples	Prop.	# Samples	Prop.
General						
COCO Caption	391219	34.97%	12828	13.83%	54707	17.21%
Vision FLAN	186060	16.63%	11922	12.85%	50068	15.75%
Image Textualization	99573	8.90%	1523	1.64%	7760	2.44%
Cambrian (filtered)	83125	7.43%	12997	14.01%	39517	12.43%
ShareGPT4O	57284	5.12%	4533	4.89%	15400	4.84%
ShareGPT4V (coco)	50017	4.47%	1510	1.63%	9755	3.07%
ALLaVA Inst (LAION GPT4V)	49990	4.47%	4696	5.06%	9976	3.14%
ShareGPT4V (llava)	29990	2.68%	1613	1.74%	9768	3.07%
ALLaVA Inst (Vision FLAN)	19990	1.79%	2577	2.78%	9220	2.90%
LLaVAR	19790	1.77%	5231	5.64%	19509	6.14%
ST-VQA	17242	1.54%	2096	2.26%	7013	2.21%
AOKVQA	16534	1.48%	2188	2.36%	7486	2.35%
Visual7W	14361	1.28%	1478	1.59%	4802	1.51%
WebSight	9995	0.89%	2632	2.84%	8742	2.75%
VisText	9964	0.89%	1769	1.91%	6363	2.00%
TallyQA	9868	0.88%	1126	1.21%	7309	2.30%
ShareGPT4V (sam)	8990	0.80%	1862	2.01%	8451	2.66%
Hateful Memes	8495	0.76%	2765	2.98%	8495	2.67%
LAION GPT4V	8048	0.72%	1525	1.64%	7139	2.25%
LLaVA Pretrain LCS	6989	0.62%	1512	1.63%	6580	2.07%
VizWiz	6604	0.59%	2809	3.03%	5220	1.64%
ScienceQA	5932	0.53%	3388	3.65%	5930	1.87%
IconQA	2496	0.22%	2214	2.39%	2496	0.79%
ShareGPT4V (knowledge)	1988	0.18%	1770	1.91%	1988	0.63%
ShareGPT4V	1926	0.17%	1911	2.06%	1926	0.61%
InterGPS	1275	0.11%	1275	1.37%	1275	0.40%
CLEVR	700	0.06%	700	0.75%	700	0.22%
VQARAD	308	0.03%	308	0.33%	308	0.10%
Doc/Chart/Screen						
UReader QA	252954	36.96%	5962	7.27%	21233	10.51%
UReader Caption	91434	13.36%	1784	2.18%	6861	3.40%
RoBUT WikiSQL	74984	10.95%	5688	6.94%	20290	10.05%
RoBUT WTQ	38241	5.59%	5622	6.86%	15094	7.47%
UReader KG	37550	5.49%	5872	7.16%	21335	10.56%
ChartQA	36577	5.34%	6669	8.14%	18550	9.18%
Chart2Text	26956	3.94%	3403	4.15%	8751	4.33%
DVQA	22000	3.21%	5489	6.70%	17219	8.52%
UReader IE	17322	2.53%	1060	1.29%	3307	1.64%
Screen2Words	15725	2.30%	3256	3.97%	9225	4.57%
AI2D (InternVL)	12403	1.81%	1530	1.87%	6715	3.32%
DocVQA	10194	1.49%	1999	2.44%	5272	2.61%
RoBUT SQA	8509	1.24%	6148	7.50%	8509	4.21%
Infographic VQA	8489	1.24%	7233	8.82%	8489	4.20%
MultiHiertt	7614	1.11%	2855	3.48%	7614	3.77%
AI2D (GPT4V Detailed Caption)	4864	0.71%	2746	3.35%	4864	2.41%
AI2D (Original)	3247	0.47%	1457	1.78%	3247	1.61%
VisualMRC	3022	0.44%	3021	3.69%	3022	1.50%
HiTab	2495	0.36%	2495	3.04%	2495	1.24%
AI2D (cauldron)	2429	0.35%	1499	1.83%	2429	1.20%
VSR	2152	0.31%	1062	1.30%	2152	1.07%
FigureQA	1880	0.27%	1880	2.29%	1880	0.93%
LRV Chart	1776	0.26%	1776	2.17%	1776	0.88%
TQA	1366	0.20%	1177	1.44%	1366	0.68%
Diagram Image2Text	295	0.04%	295	0.36%	295	0.15%
Math/Reasoning						
MAVIS Manual Collection	99990	15.60%	4033	5.22%	15763	7.08%
MAVIS Data Engine	87348	13.63%	9964	12.90%	30124	13.52%
Visual Genome	86417	13.49%	2895	3.75%	14992	6.73%
GQA	72140	11.26%	3771	4.88%	13056	5.86%

Continued on next page

Continued from previous page

Source	LLaVA-OVSI		SSR-10%		SSR-30%	
	# Samples	Prop.	# Samples	Prop.	# Samples	Prop.
Geo170K QA	67823	10.58%	2330	3.02%	10150	4.56%
Geo170k Align	60242	9.40%	5067	6.56%	12398	5.56%
TabMWP	45169	7.05%	10516	13.61%	28677	12.87%
MapQA (MathV360K)	42637	6.65%	7735	10.01%	21827	9.80%
GeoQA+ (MathV360K)	17162	2.68%	3578	4.63%	16106	7.23%
UniGeo	11949	1.86%	3855	4.99%	11947	5.36%
LRV Normal	10489	1.64%	7481	9.68%	10489	4.71%
Geometry3K	9724	1.52%	3415	4.42%	9724	4.36%
GEOmVerse (MathV360K)	9298	1.45%	2326	3.01%	9029	4.05%
Super-CLEVR	8642	1.35%	2883	3.73%	6774	3.04%
CLEVR Math	5280	0.82%	2248	2.91%	5248	2.36%
RAVEN	2100	0.33%	2100	2.72%	2100	0.94%
Geo3k	2091	0.33%	1165	1.51%	2091	0.94%
PMC-VQA	1798	0.28%	1387	1.80%	1797	0.81%
GEOS	498	0.08%	498	0.64%	498	0.22%
General OCR						
OCR-VQA	80000	39.29%	2942	20.29%	15556	29.56%
Synthdog-EN	40093	19.69%	2006	13.84%	8730	16.59%
TextOCR-GPT4V	25104	12.33%	1528	10.54%	8013	15.23%
TextCaps	21942	10.78%	1520	10.48%	7262	13.80%
Rendered Text	9995	4.91%	1881	12.97%	2796	5.31%
ChromeWriting	8825	4.33%	1813	12.50%	3277	6.23%
HME100K	7450	3.66%	761	5.25%	3364	6.39%
IAM	5658	2.78%	1167	8.05%	2151	4.09%
K12 Printing	2567	1.26%	710	4.90%	982	1.87%
IIIT5K	1990	0.98%	171	1.18%	486	0.92%

Table 7: Number of samples and proportions of sources for LLaVA-OVSI and SSR selected subsets.

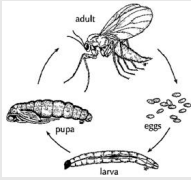
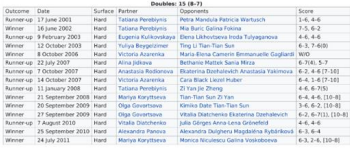
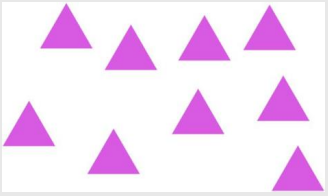
activity	Image	QA Pairs	SSR
GPT4o			
Score: 5 Style: detailed description	 Explanation: The scene is well understood with details about the environment and activities.	Q: Can you describe the main features of this image for me? A: In the heart of a bustling skatepark, a daring skateboarder is captured in mid-air, executing a thrilling trick on a wooden ramp. The skateboarder, dressed in a white shirt and beige pants, skillfully flips his skateboard beneath his feet, demonstrating his mastery and control. The ramp, the epicenter of the action, is situated in the middle of the frame. The atmosphere is vibrant and energetic. Several onlookers, some of whom are fellow skateboarders, ...	Score: 5 Style: detailed description
Score: 4 Style: detailed description	 Explanation: Activities like gathering and holding signs are recognized.	Q: Describe this image in detail with GPT-T-COCO format. A: The image captures a bustling city plaza filled with a diverse crowd of people. The crowd appears to be gathered for a demonstration, as suggested by the presence of signs and flags. The atmosphere is active and lively, with people standing, socializing, and moving around. Two people in the crowd are holding umbrellas, one of which is white. One person stands out as they are holding the umbrella in front of their face, creating an interesting visual effect. The other umbrella is positioned to the side of the image ...	Score: 4 Style: detailed description, specified style
Score: 3 Style: multi-choice	 Explanation: Recognizing the transformation between stages is key.	Q: During which stage in its life cycle is the mosquito able to lay eggs? Choices: A. Adult B. pupa C. eggs D. larva Answer with the option's letter from the given choices directly. A: A Q: Which stage transforms into adult mosquito? Choices: A. Adult B. Larva C. Pupa D. Eggs Answer with the option's letter from the given choices directly. A: C	Score: 2 Style: multi-choice
Score: 2 Style: word/short phrase, short description	 Explanation: Minimal activity recognition is required.	Q: Why is the white lines on the glass? Answer the question using a single word or phrase. A: visibility. Rationale: The white lines are to increase visibility on the translucent glass preventing people or birds from colliding into it.	Score: 2 Style: word/short phrase
Score: 1 Style: yes/no, word/short phrase	 Explanation: No activities are depicted in the image.	Q: zi yan jie zheng be the opponent on 11 january , 2008, Yes or No? Answer the question using a single word or phrase. A: Yes	Score: 0 Style: yes/no, word/short phrase
Score: 0 Style: multi-choice	 Explanation: No activity is depicted.	Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Q: How many triangles are there? Choices: (A) 9 (B) 2 (C) 10 (D) 3 (E) 4 A: The answer is A	Score: 0 Style: multi-choice

Figure 15: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **activity recognition** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

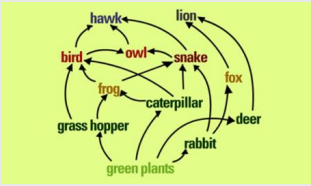
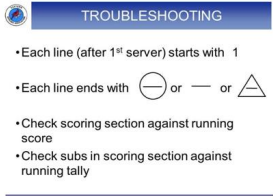
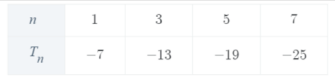
causal	Image	QA Pairs	SSR
GPT4o Score: 5 Style: word/sh ort- phrase	 Explanation: Strong causal reasoning is required to understand the impact of lions dying.	Q: What would happen if all the lions died? Less deer Fewer owls More deer Caterpillars would be extinct Answer with the given options directly. A: More deer	Score: 5 Style: word/sh ort- phrase
Score: 4 Style: detailed descripti on	 Explanation: Good causal reasoning about climate and elevation.	Q: Based on the vegetation and the snow conditions observed, what can be inferred about the likely elevation and climate conditions of this mountainous region? A: The presence of coniferous trees, which are typically found in subalpine and alpine zones, along with the extensive snow cover suggests that this region is at a high elevation, likely above the tree line where conditions are too harsh for most trees to survive. The snow appears to be fresh, indicating recent snowfall and possibly a cold ...	Score: 4 Style: detailed descript ion
Score: 3 Style: detailed descripti on	 Explanation: Understanding cause-effect in the troubleshooting process.	Q: Based on the image, what can be inferred about the nature of the troubleshooting process outlined in the slide, and how might it relate to the tasks performed by the Eastern Tennis Association? A: The troubleshooting process seems to be related to a scoring or record-keeping system, possibly for tennis matches, as the Eastern Tennis Association is involved with tennis. The first bullet point suggests a sequence that starts with "1" after the first server...	Score: 2 Style: detailed descript ion
Score: 2 Style: word/sh ort- phrase, chain- of- thought	 Explanation: Some relevance in understanding the cause of the sequence pattern.	Q: According to the question shown in the image, please first conduct reasoning, and then answer the question and provide the final value, e.g., The answer is xxx Question: Based on the image, what type of sequence is represented in the table? A: Review the differences between consecutive terms presented in the image to determine the type of sequence. Given the constant differences, the sequence is an arithmetic sequence. So the answer is arithmetic sequence The answer is arithmetic sequence.	Score: 3 Style: word/sh ort- phrase, chain- of- thought
Score: 1 Style: short descripti on	 Explanation: Minimal cause-effect information is present.	Q: Tell me about the visual elements in this screen capture. A: Setup page of a hearing aid app.	Score: 1 Style: short descript ion
Score: 0 Style: word/sh ort- phrase	 Explanation: No causal reasoning is involved.	Q: What is the heading of the document? Answer the question using a single word or phrase. A: Executive Summary	Score: 0 Style: word/sh ort- phrase

Figure 16: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **causal reasoning** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

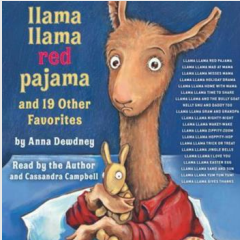
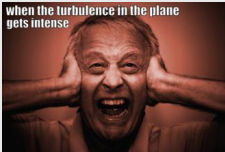
humanities	Image	QA Pairs	SSR
GPT4o			
Score: 5 Style: detailed description	 <i>Explanation: Strong understanding of mythology and cultural context is demonstrated.</i>	Q: What is the role of the solitary figure with the pole in the dynamics of this scene? A: Based on the image and the title, the solitary figure with the pole is most likely Charon, the ferryman of Hades, who in Greek and Roman mythology, transports the souls of the newly deceased across the rivers Styx and Acheron that divided the world of the living from the world of the dead. His authoritative stance and detachment from the chaos suggest he is overseeing the embarkation of the souls to their final destination in the afterlife, indifferent to their pleas and suffering.	Score: 5 Style: detailed description
Score: 4 Style: short description	 <i>Explanation: The historical nature of the engraving is relevant to humanities.</i>	Q: Can you describe the visual appearance of this image? A: Early 16th-century engraving of Toulouse	Score: 4 Style: short description
Score: 3 Style: word/short phrase, short description	 <i>Explanation: Involves literature, contributing to humanities understanding.</i>	Q: What is the title of the book mentioned in the image? A: The title of the book is "Llama Llama Red Pajama and 19 Other Favorites" by Anna Dewdney. Q: Who are the narrators of the book? A: The book is read by the author, Anna Dewdney, and Cassandra Campbell. Q: Can you list some stories or titles mentioned in the image? A: Some of the titles mentioned are: Llama Llama Red Pajama, Llama Llama Mad at Mama, Llama Llama Misses Mama, Llama Llama Holiday Drama, ...	Score: 3 Style: word/short phrase, short description
Score: 2 Style: detailed description	 <i>Explanation: Some cultural context related to media events is present.</i>	Q: Considering the consistent background and the title indicating a cast interview, what can be inferred about the nature of the event where these photos were taken, and how do the individuals' attire and expressions contribute to the setting's informality or formality? A: The nature of the event can be inferred to be a promotional interview for the television show or movie "SERVANT." The consistency in the background across all six photos suggests they were taken at the same venue, likely set up for media interviews...	Score: 2 Style: detailed description
Score: 1 Style: yes/no, word/short phrase	 <i>Explanation: Minimal cultural context is provided by the meme format.</i>	Q: Can this meme be considered disrespectful? Answer the question using a single word or phrase. A: No.	Score: 1 Style: yes/no, word/short phrase
Score: 0 Style: detailed description	 <i>Explanation: No humanities-related content is present.</i>	Q: What's happening in the scene? A: In the image, a man clad in a vibrant red and black motorcycle suit stands next to a matching red and black motorcycle, prominently displayed in the center of the image. The motorcycle, leaning on its side stand, is parked in front of a red garage door adorned with multiple windows. The man, holding a helmet in his left hand, appears to be inspecting the motorcycle with a keen eye, standing very close to the camera...	Score: 0 Style: detailed description

Figure 17: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **humanities** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

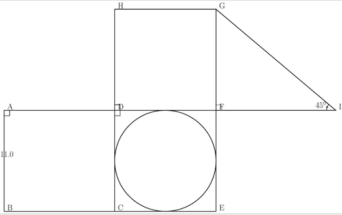
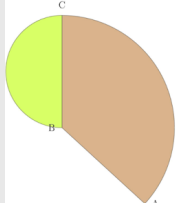
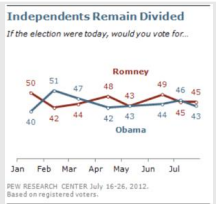
STEM	Image	QA Pairs	SSR
GPT4o		Q: According to the question shown in the image, please first perform reasoning, then finally select the right answer from the choices, e.g., Answer: xxx. Question: How would you calculate the perimeter of GFI? Choices: A: 54.5 B: 46.5 C: 52.0 D: 41.03 A: Since AB is 11.0 and ABCD is a rectangle, the opposite sides are the same length. Therefore, DC is also 11.0. DC equals 11.0. As DCEF is a square, all its sides are of equal length. Thus, DF = DC = 11.0. DF equals 11.0. As DFGH is a square, all its sides are of equal length. Thus, GF = DF = 11.0. With GF = 11.0 and angle I being 40°, ...	Score: 5 Style: multi-choice, chain-of-thought
Score: 4		Q: If the arc length of the ABC sector is 23.13 and the circumference of the lime semi-circle is 25.7, compute the degree of the CBA angle. Assume $\pi=3.14$. Round computations to 2 decimal places. A: The circumference of the lime semi-circle is 25.7 so the BC diameter can be computed as $\frac{25.7}{1 + \frac{3.14}{2}} = \frac{25.7}{2.57} = 10$. The BC radius of the ABC sector is 10 and the arc length is 23.13. So the CBA angle can be computed as $\frac{\text{ArcLength}}{2 \pi r} * 360 = \frac{23.13}{2 \pi} * 360 = \dots$	Score: 5 Style: detailed description
Score: 3		Q: Are these items generally considered rose-like in their odor? Answer the question using a single word or phrase. A: No Q: What vitamin is the vegetable particular high in? A: Vitamin c Q: What color is the food? A: Green	Score: 2 Style: yes/no, word/short-phrase
Score: 2		Q: Can you break down the data visualization and explain its message? A: The battle for independent voters remains tight. The current survey finds that 45% of independents back Romney and 43% Obama, which is virtually unchanged from earlier in July. Over the course of the year, independent support has wavered, with neither candidate holding a consistent advantage.	Score: 2 Style: detailed description
Score: 1		Q: What precautions should the man take while working in the rain? Answer the question with GPT-T-COCO format. A: While working in the rain, the man should take several precautions to ensure his safety and protect the items he handles. These precautions include wearing appropriate rain gear such as waterproof clothing, a raincoat, and footwear with slip-resistant soles to avoid potential slips and falls caused by wet surfaces. Additionally, he should exercise extra ...	Score: 1 Style: detailed description
Score: 0		Q: Is the bus moving? Answer the question using a single word or phrase. A: Yes Q: Is this a movie studio backlot? A: No Q: What is the primary color of the bus? A: Yellow Q: Is there are place of worship in the photo? A: No	Score: 0 Style: yes/no, word/short-phrase

Figure 18: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **stem knowledge** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

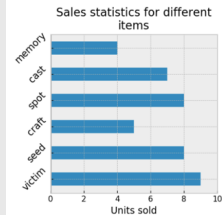
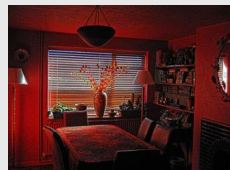
comparative	Image	QA Pairs	SSR																																																												
GPT4o Score: 5 Style: word/short-phrase, yes/no, comparison	 Explanation: Strong need for comparing quantities across items.	Q: Which item sold the most units? Answer the question with a single word. A: Victim Q: Which item sold the least units? A: Memory Q: How many units of the the most sold item were sold? A: 9 Q: How many units of the the least sold item were sold? A: 4 ...	Score: 4 Style: word/short-phrase, yes/no, comparison																																																												
Score: 4 Style: detailed description, comparison	 Explanation: The comparison between messy and tidy states is clear.	Q: Describe this image in detail with GPT-T-COCO format. A: The image presents a before-and-after comparison of a room undergoing a transformation from a messy state to a clean and tidy one. In the 'before' part of the image, the room is quite cluttered. There are books scattered about, a disorganized desk with a laptop on it, and a chair pushed askew. A suitcase is also visible, contributing to the sense of disarray. The 'after' part of the image showcases the same room, but with a considerable difference. It's noticeably tidier, with the books neatly stacked, the chair properly adjusted, and the laptop closed. The suitcase appears to be packed away, and the overall space looks much more organized...	Score: 4 Style: detailed description, comparison																																																												
Score: 3 Style: yes/no, word/short-phrase	<table><tr><th colspan="5">Medal table</th></tr><tr><th>Nation</th><th>Gold</th><th>Silver</th><th>Bronze</th><th>Total</th></tr><tr><td>Cyprus</td><td>32</td><td>30</td><td>20</td><td>82</td></tr><tr><td>Luxembourg</td><td>30</td><td>15</td><td>27</td><td>72</td></tr><tr><td>Iceland</td><td>20</td><td>23</td><td>25</td><td>68</td></tr><tr><td>Malta</td><td>8</td><td>12</td><td>9</td><td>29</td></tr><tr><td>Liechtenstein</td><td>6</td><td>10</td><td>11</td><td>27</td></tr><tr><td>Monaco</td><td>6</td><td>9</td><td>14</td><td>29</td></tr><tr><td>Montenegro</td><td>4</td><td>2</td><td>2</td><td>8</td></tr><tr><td>Andorra</td><td>3</td><td>7</td><td>5</td><td>15</td></tr><tr><td>San Marino</td><td>3</td><td>4</td><td>11</td><td>18</td></tr><tr><td>Total</td><td>112</td><td>112</td><td>124</td><td>348</td></tr></table> Explanation: Involves comparing medal counts.	Medal table					Nation	Gold	Silver	Bronze	Total	Cyprus	32	30	20	82	Luxembourg	30	15	27	72	Iceland	20	23	25	68	Malta	8	12	9	29	Liechtenstein	6	10	11	27	Monaco	6	9	14	29	Montenegro	4	2	2	8	Andorra	3	7	5	15	San Marino	3	4	11	18	Total	112	112	124	348	Q: cyprus win the most gold medal with 32, Yes or No? Answer the question using a single word or phrase. A: Yes	Score: 3 Style: yes/no, word/short-phrase
Medal table																																																															
Nation	Gold	Silver	Bronze	Total																																																											
Cyprus	32	30	20	82																																																											
Luxembourg	30	15	27	72																																																											
Iceland	20	23	25	68																																																											
Malta	8	12	9	29																																																											
Liechtenstein	6	10	11	27																																																											
Monaco	6	9	14	29																																																											
Montenegro	4	2	2	8																																																											
Andorra	3	7	5	15																																																											
San Marino	3	4	11	18																																																											
Total	112	112	124	348																																																											
Score: 2 Style: detailed description, specified style	 Explanation: Some comparison is implied between the sandwiches.	Q: Describe this image in detail with GPT-T-COCO format. A: The image depicts a restaurant scene where a large, appetizing sandwich is the center of attention. The sandwich, resting on a large plate, is piled high with rare roast beef, suggesting a corned beef sandwich. It's an open sandwich, with the slices of bread and layers of meat clearly visible... A second sandwich can be glimpsed further on the dining ...	Score: 1 Style: detailed description, specified style																																																												
Score: 1 Style: short description	 Explanation: No comparison is required or possible.	Q: Provide a one-sentence caption for the provided image. A: A dining room bathed in red light with a floral arrangement.	Score: 1 Style: short description																																																												
Score: 0 Style: word/short-phrase	 Explanation: No comparison is required.	Q: Transcribe visible text from this photograph. A: rendered in the Greek Revival style.	Score: 0 Style: word/short-phrase																																																												

Figure 19: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **comparative analysis** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

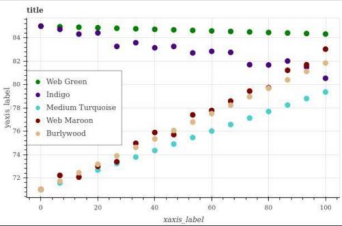
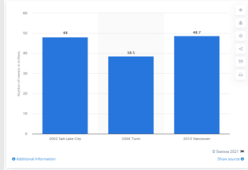
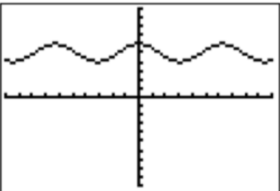
data	Image	QA Pairs	SSR
GPT4o			
Score: 5 Style: detailed description	 Explanation: Strong understanding of data representation in scatter plots.	Q: Write a detailed and comprehensive description of the image. A: The image is a scatter plot with five distinct categories represented by different colored markers. Each category is plotted against clearly defined x and y axes. ### Graph Elements: - **Title**: The plot has a placeholder for the title labeled "title". - **X-axis Label**: The x-axis is labeled "xaxis label". - **Y-axis Label**: The y-axis is labeled "yaxis label". ### Data Points: - **Web Green**: Represented by green circle markers. These points are clustered at the top of the plot around y values of approximately 82 to 86...	Score: 5 Style: detailed description
Score: 4 Style: word/sh ort-phrase	 Explanation: Significant as it involves interpreting a chart to extract information.	Q: What was the average minute rating of Olympic Winter Games broadcasts in Salt Lake City in 2002? Answer the question using a single word or phrase. A: 48.	Score: 3 Style: word/sh ort-phrase
Score: 3 Style: multi-choice, chain-of-thought	 Explanation: Involves knowledge of vitamins in vegetables.	Q: According to the question shown in the image, please first perform reasoning, then finally select the right answer from the choices, e.g., Answer: xxx. Question: Based on the image, what is the minimum value of the cosine function? Choices: A. -1 B. 0 C. 4 D. 5 A: The minimum value of the cosine function can be identified as the lowest point of its cycle. With an amplitude of 1 and a vertical shift of 5, the minimum value is $5 - 1 = 4$. So the answer is C Answer: C	Score: 3 Style: multi-choice, chain-of-thought
Score: 2 Style: short description	 Explanation: The image includes structured text, useful for data understanding.	Q: Provide a one-sentence caption for the provided image. A: A translated page that is titled Attacking Ramsus' vessel!!	Score: 2 Style: short description
Score: 1 Style: coordinate, short description	 Explanation: Basic data understanding in terms of object identification.	Q: Provide a short description for the given region. [0.658,0.832,0.816,0.925] A: A blue and white truck. Q: [0.738,0.853,0.964,0.963] A: A silver car. Q: [0.280,0.840,0.456,0.939] A: A white house. Q: [0.138,0.915,0.172,0.973] ...	Score: 0 Style: coordinate, short description
Score: 0 Style: short description, word/sh ort-phrase	 Explanation: No data or charts to interpret.	Q: What color is the handle of the knife? Answer the question using a single word or phrase. A: Green Q: What is the man doing? A: Cutting apple Q: What type of apple is he holding? A: Red	Score: 0 Style: short description, word/sh ort-phrase

Figure 20: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **data understanding** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

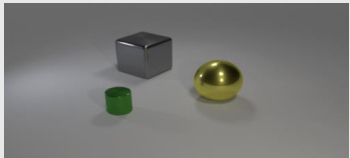
spatial	Image	QA Pairs	SSR
GPT4o Score: 5 Style: coordina te	 Explanation: The task requires identifying precise locations of objects, enhancing spatial understanding.	Q: Provide the bounding box coordinates of the region that the given sentence describes. The passing vehicle. A: [0.000,0.000,0.662,0.757] Q: A large black and dark green backpack on the ground. A: [0.253,0.465,0.500,0.991] Q: The bus that is behind the luggage and the person on the sidewalk. it is in motion. A: [0.000,0.000,0.662,0.757] ...	Score: 5 Style: coordina ate
Score: 4 Style: multi- choice, yes/no	 Explanation: The image requires understanding of symmetry and spatial alignment.	Q: Is the dotted line a line of symmetry? Choices: A. yes B. no Answer with the option's letter from the given choices directly. A: A Q: Does this picture have symmetry? Choices: A. no B. yes Answer with the option's letter from the given choices directly. A: B	Score: 4 Style: multi- choice, yes/no
Score: 3 Style: detailed descripti on	 Explanation: The description includes spatial relationships, aiding understanding.	Q: Can you describe the lighting in the scene? A: The lighting in the scene appears to be diffused, with a soft shadow cast beneath each object, indicating an overhead light source that is not too harsh. There are no strong highlights, which suggests the lighting is even and possibly from a studio light setup or image rendering with a global illumination feature for realistic lighting.	Score: 3 Style: detailed descript ion
Score: 2 Style: detailed descripti on	 Explanation: Some spatial understanding is possible, but <i>the description is inaccurate</i>.	Q: What's happening in the scene? A: The image features a white plate holding a delicious meal, consisting of a <i>half-eaten sandwich</i> on a roll, which occupies most of the space, and a side salad placed near the front left corner of the plate. The salad is made up of various mixed vegetables such as broccoli, cauliflower, cabbage, and other greens. In the front of the plate, closer to the camera, a glazed tan round <i>donut hole is positioned</i>. The arrangement creates a visually appealing and appetizing dish.	Score: 2 Style: detailed descript ion
Score: 1 Style: yes/no, word/sh ort- phrase	 Explanation: No spatial relationship with a dark car is visible.	Q: Is the hydrant to the left or to the right of the dark car? Answer the question using a single word or phrase. A: Left Q: What is common to the sign and the fire hydrant? A: Color Q: Which place is it? A: Road Q: Is there any black signal light or fire hydrant? A: No	Score: 4 Style: yes/no, word/sh ort- phrase
Score: 0 Style: detailed descripti on	 Explanation: No spatial relationships or object positions are relevant in this context.	Q: Considering the specific project title listed under "CURRENT FUNDING OPPORTUNITIES", what could be the focus of research for Early Stage Researchers in the context of "Tendon Repair, Network of Excellence for Functional Biomaterials"? A: Based on the project title, the focus of research for Early Stage Researchers is likely to be on developing or improving biomaterials that can be used in the repair of tendons. The "Network of Excellence" suggests a collaborative effort, potentially involving multiple institutions or departments, aimed at innovating in the field of functional biomaterials specifically for tendon repair applications.	Score: 0 Style: detailed descript ion

Figure 21: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **object spatial understanding** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

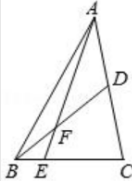
attribute	Image	QA Pairs	SSR																																																																																						
GPT4o Score: 5 Style: detailed description, chain-of-thought	 Explanation: Detailed attributes of the watch are identified, including brand, color, and features.	Q: Can you describe all the objects and characters in the picture? A: The image depicts a wristwatch with a prominent and detail-rich dial, set against a green backdrop. To the right bottom corner, a part of a date-stamped paper is visible. The watch is contained in a transparent plastic protective case or wrapper. Here is a detailed description and analysis: ### Watch Dial: - **Brand and Model**: The brand name "Audemars Piguet" can be read on the dial, indicating this is a luxury timepiece highly regarded in ...	Score: 5 Style: detailed description, chain-of-thought																																																																																						
Score: 4 Style: multi-choice, yes/no	 Explanation: Attributes like the cow's color and the man's hat are correctly identified.	Q: Why are the cows wearing ribbons? Answer the question using a single word or phrase. A: 1 Q: Is the man wearing a hat? A: Yes Q: What number is on the cow? A: 3010 Q: How many different types of head coverings are people wearing? A: 1 Q: Does this cow have black and white spots? A: Yes Q: How tall is the cow? ...	Score: 4 Style: multi-choice, yes/no																																																																																						
Score: 3 Style: short description	 Explanation: Identifying the computer and its screen provides moderate attribute information.	Q: Render a clear and concise summary of the photo. A: Procomm Plus written in blue on a computer	Score: 3 Style: short description																																																																																						
Score: 2 Style: yes/no, word/short-phrase	<table><tr><th>Week</th><th>Date</th><th>Schedule</th><th>Opponent</th><th>Result</th><th>Attendance</th></tr><tr><td>1</td><td>September 3, 1978</td><td>at New Orleans Saints</td><td>L 31-24</td><td>54,187</td></tr><tr><td>2</td><td>September 11, 1978</td><td>Denver Broncos</td><td>W 12-9</td><td>46,508</td></tr><tr><td>3</td><td>September 17, 1978</td><td>Tampa Bay Buccaneers</td><td>L 16-10</td><td>46,132</td></tr><tr><td>4</td><td>September 25, 1978</td><td>at Chicago Bears</td><td>W 24-20</td><td>53,551</td></tr><tr><td>5</td><td>October 1, 1978</td><td>at Tampa Bay Buccaneers</td><td>W 24-7</td><td>65,972</td></tr><tr><td>6</td><td>October 8, 1978</td><td>at Seattle Seahawks</td><td>L 29-28</td><td>62,031</td></tr><tr><td>7</td><td>October 15, 1978</td><td>Los Angeles Rams</td><td>L 34-17</td><td>46,551</td></tr><tr><td>8</td><td>October 22, 1978</td><td>Green Bay Packers</td><td>W 21-7</td><td>47,411</td></tr><tr><td>9</td><td>October 26, 1978</td><td>at Dallas Cowboys</td><td>W 21-10</td><td>61,848</td></tr><tr><td>10</td><td>November 5, 1978</td><td>Detroit Lions</td><td>W 17-7</td><td>46,008</td></tr><tr><td>11</td><td>November 12, 1978</td><td>Chicago Bears</td><td>W 17-14</td><td>43,286</td></tr><tr><td>12</td><td>November 19, 1978</td><td>San Diego Chargers</td><td>L 13-7</td><td>38,859</td></tr><tr><td>13</td><td>November 26, 1978</td><td>at Green Bay Packers</td><td>T 10-10 (OT)</td><td>51,737</td></tr><tr><td>14</td><td>December 3, 1978</td><td>Philadelphia Eagles</td><td>W 28-21</td><td>38,723</td></tr><tr><td>15</td><td>December 9, 1978</td><td>at Detroit Lions</td><td>L 45-14</td><td>78,685</td></tr><tr><td>16</td><td>December 17, 1978</td><td>at Oakland Raiders</td><td>L 27-20</td><td>44,643</td></tr></table> Explanation: Some identification of game results is needed, but it's minimal.	Week	Date	Schedule	Opponent	Result	Attendance	1	September 3, 1978	at New Orleans Saints	L 31-24	54,187	2	September 11, 1978	Denver Broncos	W 12-9	46,508	3	September 17, 1978	Tampa Bay Buccaneers	L 16-10	46,132	4	September 25, 1978	at Chicago Bears	W 24-20	53,551	5	October 1, 1978	at Tampa Bay Buccaneers	W 24-7	65,972	6	October 8, 1978	at Seattle Seahawks	L 29-28	62,031	7	October 15, 1978	Los Angeles Rams	L 34-17	46,551	8	October 22, 1978	Green Bay Packers	W 21-7	47,411	9	October 26, 1978	at Dallas Cowboys	W 21-10	61,848	10	November 5, 1978	Detroit Lions	W 17-7	46,008	11	November 12, 1978	Chicago Bears	W 17-14	43,286	12	November 19, 1978	San Diego Chargers	L 13-7	38,859	13	November 26, 1978	at Green Bay Packers	T 10-10 (OT)	51,737	14	December 3, 1978	Philadelphia Eagles	W 28-21	38,723	15	December 9, 1978	at Detroit Lions	L 45-14	78,685	16	December 17, 1978	at Oakland Raiders	L 27-20	44,643	Q: they lose the first game to the new orleans saint the final score be 31 - 24, Yes or No? Answer the question using a single word or phrase. A: Yes	Score: 2 Style: yes/no, word/short-phrase
Week	Date	Schedule	Opponent	Result	Attendance																																																																																				
1	September 3, 1978	at New Orleans Saints	L 31-24	54,187																																																																																					
2	September 11, 1978	Denver Broncos	W 12-9	46,508																																																																																					
3	September 17, 1978	Tampa Bay Buccaneers	L 16-10	46,132																																																																																					
4	September 25, 1978	at Chicago Bears	W 24-20	53,551																																																																																					
5	October 1, 1978	at Tampa Bay Buccaneers	W 24-7	65,972																																																																																					
6	October 8, 1978	at Seattle Seahawks	L 29-28	62,031																																																																																					
7	October 15, 1978	Los Angeles Rams	L 34-17	46,551																																																																																					
8	October 22, 1978	Green Bay Packers	W 21-7	47,411																																																																																					
9	October 26, 1978	at Dallas Cowboys	W 21-10	61,848																																																																																					
10	November 5, 1978	Detroit Lions	W 17-7	46,008																																																																																					
11	November 12, 1978	Chicago Bears	W 17-14	43,286																																																																																					
12	November 19, 1978	San Diego Chargers	L 13-7	38,859																																																																																					
13	November 26, 1978	at Green Bay Packers	T 10-10 (OT)	51,737																																																																																					
14	December 3, 1978	Philadelphia Eagles	W 28-21	38,723																																																																																					
15	December 9, 1978	at Detroit Lions	L 45-14	78,685																																																																																					
16	December 17, 1978	at Oakland Raiders	L 27-20	44,643																																																																																					
Score: 1 Style: detailed description	 Explanation: Minimal attributes are identified in the description.	Q: Describe what you see in the figure. A: In the diagram, there is a triangle labeled ABC, with points A, B, and C as its vertices. On the line segment BC, there's a point named E. Point D is the midpoint of the line segment AC.	Score: 2 Style: detailed description																																																																																						
Score: 0 Style: detailed description	 Explanation: No attributes of objects are identified in this instance.	Q: What might be the implications of the viewpoint expressed in the tweet regarding the use of encryption in consumer products, and how could this stance affect the privacy and security landscape if implemented as a policy? A: Based on the tweet, the user is expressing a stance against the use of encryption for routine communications in consumer products, suggesting that security should be ...	Score: 0 Style: detailed description																																																																																						

Figure 22: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **attribute identification** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

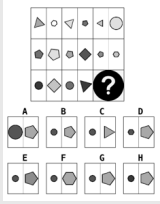
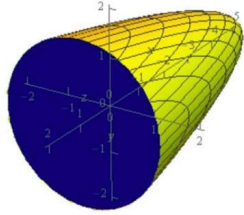
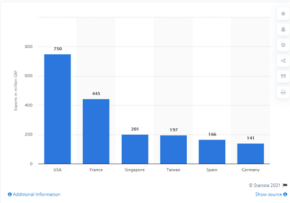
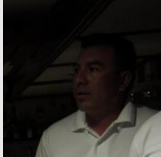
logical	Image	QA Pairs	SSR
GPT4o			
Score: 5 Style: multi-choice	 Explanation: Logical reasoning is key to determining the sequence completion.	Q: Which figure should complete the logical sequence? A: G	Score: 5 Style: multi-choice
Score: 4 Style: word/sh ort- phrase, chain- of- thought	 Explanation: Requires reasoning to deduce the shape from given information.	Q: According to the question shown in the image, please first conduct reasoning, and then answer the question and provide the final value, e.g., The answer is xxx Question: Based on the image, what is the shape of the region D where the double integration is performed? A: Region D where the double integration takes place is shown as a circular disk in the yz-plane. The limits for this region are derived from the intersection of the paraboloid and the plane, yielding a circle of radius $\sqrt{3}$, which is highlighted in the diagram. So the answer is circular disk.	Score: 4 Style: word/sh ort- phrase, chain- of- thought
Score: 3 Style: detailed descripti on	 Explanation: Some deduction about time and season is present.	Q: What time of day and season could this painting be representing based on the lighting and the activities of the people? A: The painting could be representing late morning to early afternoon, as the light is bright but diffused, suggesting the sun is not at its peak. The presence of clouds might be blocking the direct sunlight, casting a soft glow over the landscape. The season appears to be either late spring or early autumn, as the fields are lush but not in the height of summer growth, ...	Score: 3 Style: detailed descript ion
Score: 2 Style: specifie d style	 Explanation: Limited logical connections are made between elements.	Q: You are provided with a picture, write a caption with a specific sentiment (positive or negative) related to the picture. Note that the sentiment in the caption should match the requested sentiment. Write a caption with a negative sentiment for the given image. A: a horrible picture of people at the mountain ski resort	Score: 1 Style: specifie d style
Score: 1 Style: word/sh ort- phrase	 Explanation: Minimal logical deduction is required.	Q: What was the value share of Scotch whisky in the United States in 2014? Answer the question using a single word or phrase. A: 750.	Score: 3 Style: word/sh ort- phrase
Score: 0 Style: word/sh ort- phrase	 Explanation: The task does not require logical deduction.	Q: Here is a picture of a person. Based only upon this picture, what would you guess this person's gender is? A: The person's gender is Male	Score: 0 Style: word/sh ort- phrase

Figure 23: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **logical deduction** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

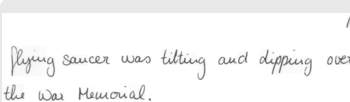
scene	Image	QA Pairs	SSR																		
GPT4o Score: 5 Style: detailed description	 Explanation: The description offers a comprehensive understanding of the scene.	Q: Describe this image in detail. A: The image captures the entrance to the City Museum in St. Louis, Missouri. Dominating the scene is a large, rusted metal gate that serves as the main entrance. The gate is adorned with a sign that spells out "City Museum" in vibrant, colorful letters. On either side of the gate, two imposing metal sculptures stand guard. One of these sculptures takes the form of a dragon, adding a touch of whimsy to the scene. The ground beneath these structures is blanketed with fallen leaves, hinting at a season change...	Score: 5 Style: detailed description																		
Score: 4 Style: detailed description	 Explanation: The scene is well described, capturing the overall ambiance and design.	Q: Describe this image in detail. A: In the midst of a snowy landscape, a pair of legs clad in vibrant orange ski boots stand confidently. The boots, adorned with black buckles, are firmly attached to a pair of brown skis streaked with white stripes. The skis cut a striking contrast against the pristine white snow, which appears freshly fallen and untouched except for a few tracks in the background. In the skier's grip are two ski poles, their black shafts ending in white handles. The poles are poised as if ready to push off into the snow at any moment...	Score: 4 Style: detailed description																		
Score: 3 Style: detailed description	 Explanation: The scene is simple but provides context for the alliance theme.	Q: Considering the context provided in the slide, what could be the significance of the symbols of the gavel and medical bag in relation to the text presented? A: The gavel and the medical bag symbols next to the title "Alliance formed" suggest that the alliance involves legal and medical aspects. This is supported by the text, which refers to characters named Armstrong and Wargrave. Armstrong is likely a medical professional, suggested by the medical bag, and Wargrave, who may be the person in the judge's robe, represents the legal side, indicated by the gavel. These symbols are visual cues to the professions or roles of the characters involved ...	Score: 3 Style: detailed description																		
Score: 2 Style: word/short-phrase	 Explanation: Minimal scene understanding is required beyond identifying the train and its text	Q: What are the innitals on the train? Answer the question using a single word or phrase. A: BNSF.	Score: 1 Style: word/short-phrase																		
Score: 1 Style: detailed description	<table><thead><tr><th>Price</th><th>Quantity demanded</th><th>Quantity supplied</th></tr></thead><tbody><tr><td>\$450</td><td>18,800</td><td>10,300</td></tr><tr><td>\$580</td><td>15,700</td><td>10,400</td></tr><tr><td>\$710</td><td>12,600</td><td>10,500</td></tr><tr><td>\$840</td><td>9,500</td><td>10,600</td></tr><tr><td>\$970</td><td>6,400</td><td>10,700</td></tr></tbody></table> Explanation: Basic understanding of a table, not a complex scene.	Price	Quantity demanded	Quantity supplied	\$450	18,800	10,300	\$580	15,700	10,400	\$710	12,600	10,500	\$840	9,500	10,600	\$970	6,400	10,700	Q: Look at the table. Then answer the question. At a price of \$710, is there a shortage or a surplus? A: At the price of \$710, the quantity demanded is greater than the quantity supplied. There is not enough of the good or service for sale at that price. So, there is a shortage.	Score: 1 Style: detailed description
Price	Quantity demanded	Quantity supplied																			
\$450	18,800	10,300																			
\$580	15,700	10,400																			
\$710	12,600	10,500																			
\$840	9,500	10,600																			
\$970	6,400	10,700																			
Score: 0 Style: word/short-phrase	 Explanation: The image does not depict a scene.	Q: What's written in this image? A: A flying saucer was tilting and dipping over the War Memorial.	Score: 0 Style: word/short-phrase																		

Figure 24: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **scene understanding** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

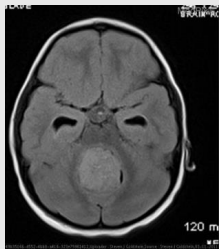
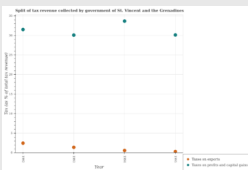
fine-grained	Image	QA Pairs	SSR
GPT4o Score: 5 Style: word/sh ort- phrase	 Explanation: High level of detail needed to distinguish aircraft variant.	Q: Your objective is to analyze an aircraft image and provide the variant of the aircraft. (e.g., A300B4). Variant: A variant indicates a variation of a particular aircraft model, often incorporating specific modifications, improvements, or customizations compared to the base model. A: BAE 146-300	Score: 5 Style: word/sh ort- phrase
Score: 4 Style: detailed descripti on	 Explanation: Specific brain structures are identified with some detail.	Q: Describe the content of a given image in detail A: This image is an MRI scan of a human brain, oriented to provide an axial view. An MRI (Magnetic Resonance Imaging) scan uses strong magnetic fields and radio waves to generate detailed images of the inside of the body, particularly the soft tissues. In this particular slice, we can observe several key structures of the brain. 1. Cerebral Hemispheres: The left and right cerebral hemispheres are clearly visible. The presence of sulci and gyri, the folds of the brain, is evident although not extremely detailed in this specific slice. 2. Corpus Callosum: The corpus callosum...	Score: 4 Style: detailed descripti on
Score: 3 Style: multi- choice	 Explanation: Recognizing specific features of the elephant is required.	Q: Considering the size of his ears what continent is this elephant from? A. north america B. asia C. africa D. europe Answer with the option's letter from the given choices directly. A: B	Score: 3 Style: multi- choice
Score: 2 Style: word/sh ort- phrase, yes/no	 Explanation: Limited detail on the plane's specific model or era.	Q: What continent is this plane flying over? Answer the question using a single word or phrase. A: Antarctica Q: What is below the plane? A: Mountains Q: Is this a plane made after the year 2000? A: No	Score: 2 Style: word/sh ort- phrase, yes/no
Score: 1 Style: compari son	 Explanation: Minimal need for distinguishing subtle differences.	Q: You are provided with a chart image and a question related to the chart. Answer the question based on the information given in the chart. In the year 1992, what is the difference between the percentage of revenue obtained from taxes on profits and capital gains and percentage of revenue obtained from taxes on exports ? A: 29.05551550108148	Score: 1 Style: compari son
Score: 0 Style: multi- choice, chain- of- thought	 Explanation: No fine-grained recognition is needed for the image.	Q: According to the question shown in the image, please first perform reasoning, then finally select the right answer from the choices, e.g., Answer: xxx. Question: Based on the image, what mathematical operation is used on side \$b\$ and the hypotenuse to find side \$a\$? Choices: A. Addition B. Subtraction C. Multiplication D. Division A: You square and then subtract the measurements to solve for side \$a\$ using the Pythagorean theorem. So the answer is B Answer: B	Score: 0 Style: multi- choice, chain- of- thought

Figure 25: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **fine-grained recognition** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

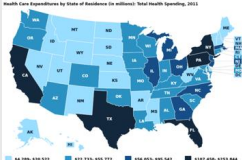
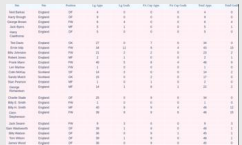
lang	Image	QA Pairs	SSR
GPT4o		Q: Analyze and describe in detail the visual elements in this image. A: The image showcases a vibrant town square that includes a well-preserved, large stone column monument at its center. This column appears to be an artistic sculpture featuring multiple figures and intricate carvings. The column is the central focus and reaches a significant height, visually dominating the scene. Atop the column, there is a prominent figure holding a cross, which is accented in gold, suggesting religious significance. Surrounding this central figure are several other figures, possibly angels or historical characters, integrated into the structure with dynamic, flowing poses. The immediate surroundings are ...	Score: 5 Style: detailed description
		Q: What do you see happening in this image? A: The image captures a moment at a train station. On the left, a vibrant yellow and blue train, with a yellow front, is approaching the viewer from a distance, moving towards the camera on the right track, as if ready to embark on a new journey. On the right side of the image, close to the camera, another train is seen. This one is adorned in red and white, with a contrasting blue front, and is moving away from the viewer on the left track, as if it has completed its journey and is now returning. The station itself, with its red roof, nestles comfortably in the background. It's a silent observer to the comings and goings of these mechanical beasts ...	Score: 4 Style: detailed description
		Q: Provide a one-sentence caption for the provided image. A: The clock shows one hand at around 10 and another at around 12.	Score: 2 Style: short description
		Q: What is the value of Idaho? Answer the question using a single word or phrase. A: 4,289-20,522. Q: Name the states that have a value in the range 107,458-253,844? A: California, Florida, New York, Pennsylvania, Texas. ...	Score: 2 Style: word/short phrase, yes/no, comparison
		Q: What are the number of league apps ted davis has? Answer the question using a single word or phrase. A: 27.	Score: 1 Style: word/short phrase
		Q: Provide the bounding box coordinates of the region that the given sentence describes. Hitter. A: [0.491,0.281,0.720,0.771] Q: Catcher. A: [0.122,0.496,0.380,0.796] ...	Score: 0 Style: coordinate

Figure 26: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **language generation** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

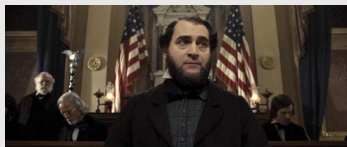
in-context	Image	QA Pairs	SSR																																																																	
GPT4o Score: 5 Style: detailed description, short description	 Explanation: The instance follows the task instructions and provides answers in the specified format.'	Q: What are the main features of the kitchen in the image? Answer the question with GPT-T-COCO format. A: The main features of the kitchen in the image include a stove with an oven, a table, various pots and utensils, and a light that is turned on next to the stove. The kitchen has a rustic appearance. Q: Is there anything placed on top of the oven? A: Yes, there is a bowl placed on top of the oven. Q: What is the overall size of the kitchen? ...	Score: 3 Style: detailed description, short description																																																																	
Score: 4 Style: detailed description, specified style	 Explanation: The speculative scenario provides a creative context for learning.	Q: Explain the significance of the main character's attire. A: The main character's attire is highly significant, as it reflects the fashion and societal norms of a specific historical period. Dressed in a black suit with a formal shirt and a meticulously groomed beard, he presents a dignified and authoritative figure. Such attire was typical for gentlemen in the 19th century, particularly those in positions of power or influence, such as lawyers, politicians, or other public figures. This choice of clothing helps to convey the character's social status and the serious nature of the proceedings he is involved in. Q: What does the presence of the American flags behind him symbolize? ...	Score: 3 Style: detailed description, specified style																																																																	
Score: 3 Style: multi-choice	 Explanation: Follows the task format and context well.	Q: What type of animal is featured in the image alongside the babies? Answer the question with GPT-T-COCO format. A: The image features a mother sheep, also known as a ewe, alongside her two baby sheep, also known as lambs. Q: What are the baby lambs doing in the image? A: In the image, one baby lamb is eating from its mother, while the other baby lamb is eating grass. ...	Score: 3 Style: multi-choice																																																																	
Score: 2 Style: word/short phrase, yes/no	<table><tr><th>Place</th><th>Player</th><th>Country</th><th>Score</th><th>To par</th></tr><tr><td>1</td><td>Bobby Clampett</td><td>United States</td><td>67-66=133</td><td>-11</td></tr><tr><td>2</td><td>Nick Price</td><td>Zimbabwe</td><td>69-69=138</td><td>-6</td></tr><tr><td>T2</td><td>Bernhard Langer</td><td>West Germany</td><td>70-69=139</td><td>-5</td></tr><tr><td></td><td>Des Smyth</td><td>Ireland</td><td>70-69=139</td><td>-5</td></tr><tr><td>T5</td><td>Sandy Lyle</td><td>Scotland</td><td>74-66=140</td><td>-4</td></tr><tr><td></td><td>Tom Watson</td><td>United States</td><td>69-71=140</td><td>-4</td></tr><tr><td>T7</td><td>Ken Brown</td><td>Scotland</td><td>70-71=141</td><td>-3</td></tr><tr><td></td><td>Peter Oosterhuis</td><td>England</td><td>74-67=141</td><td>-3</td></tr><tr><td>9</td><td>Tom Purtzer</td><td>United States</td><td>76-68=142</td><td>-2</td></tr><tr><td></td><td>José María Cañizares</td><td>Spain</td><td>71-72=143</td><td>-1</td></tr><tr><td>T10</td><td>David Graham</td><td>Australia</td><td>73-70=143</td><td>-1</td></tr><tr><td></td><td>Bill Rogers</td><td>United States</td><td>73-70=143</td><td>-1</td></tr></table> Explanation: Limited demonstration of task understanding.	Place	Player	Country	Score	To par	1	Bobby Clampett	United States	67-66=133	-11	2	Nick Price	Zimbabwe	69-69=138	-6	T2	Bernhard Langer	West Germany	70-69=139	-5		Des Smyth	Ireland	70-69=139	-5	T5	Sandy Lyle	Scotland	74-66=140	-4		Tom Watson	United States	69-71=140	-4	T7	Ken Brown	Scotland	70-71=141	-3		Peter Oosterhuis	England	74-67=141	-3	9	Tom Purtzer	United States	76-68=142	-2		José María Cañizares	Spain	71-72=143	-1	T10	David Graham	Australia	73-70=143	-1		Bill Rogers	United States	73-70=143	-1	Q: 2 athlete finish 4 below par, Yes or No? Answer the question using a single word or phrase. A: Yes	Score: 2 Style: word/short phrase, yes/no
Place	Player	Country	Score	To par																																																																
1	Bobby Clampett	United States	67-66=133	-11																																																																
2	Nick Price	Zimbabwe	69-69=138	-6																																																																
T2	Bernhard Langer	West Germany	70-69=139	-5																																																																
	Des Smyth	Ireland	70-69=139	-5																																																																
T5	Sandy Lyle	Scotland	74-66=140	-4																																																																
	Tom Watson	United States	69-71=140	-4																																																																
T7	Ken Brown	Scotland	70-71=141	-3																																																																
	Peter Oosterhuis	England	74-67=141	-3																																																																
9	Tom Purtzer	United States	76-68=142	-2																																																																
	José María Cañizares	Spain	71-72=143	-1																																																																
T10	David Graham	Australia	73-70=143	-1																																																																
	Bill Rogers	United States	73-70=143	-1																																																																
Score: 1 Style: comparison	 Explanation: Limited context is provided.	Q: Analyze the image in a comprehensive and detailed manner. A: The image features a spacious living room with a clean and modern feel. There is a large, not red but rather brown, couch placed in the room, and a comfortable chair nearby, which is a wooden chair with fabric seat cover, situated away from the camera. A dining table is situated in the center of the room, surrounded by chairs, ...	Score: 1 Style: comparison																																																																	
Score: 0 Style: word/short phrase	A. ①② B. ①④ Explanation: No context-based learning is demonstrated.	Q: 图中写了什么？请直接使用图中出现的文字回答，不用考虑上下文。只回答图中出现的文字。 A: A. ①② B. ①④	Score: 0 Style: word/short phrase																																																																	

Figure 27: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **in-context learning** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.

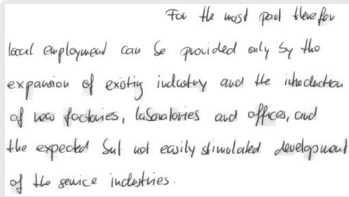
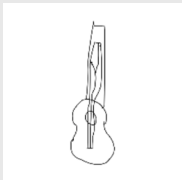
OCR	Image	QA Pairs	SSR
GPT4o			
Score: 5 Style: detailed description	 <i>For the most part therefore local employment can be provided only by the expansion of existing industry and the introduction of new factories, laboratories and offices, and the expected but not easily stimulated development of the service industries.</i> Explanation: The instance provides a clear example of handwritten text recognition, which is highly relevant for OCR capability.	Q: What is scribbled in this image? A: For the most part therefore local employment can be provided only by the expansion of existing industry and the introduction of new factories, laboratories and offices, and the expected but not easily stimulated development of the service industries.	Score: 5 Style: detailed description
Score: 4 Style: detailed description	 Explanation: The description includes details about the text on the card, aiding OCR capabilities.	Q: Describe this image in detail. A: The image displays a golden business card with intricate, laser-cut detailing around the edges. The card features a mix of English and Chinese text. In the upper portion, the Chinese characters are prominent, and below them, the English translation reads "Personal Tailor." The text is printed in red with a shimmering effect that contrasts with the metallic sheen of the gold card. The card's design is elegant and suggests a personalized service, ...	Score: 4 Style: detailed description
Score: 3 Style: short description	 Explanation: The image contains text on the truck, which can be recognized.	Q: Write a terse but informative summary of the picture. A: The yellow taxi drives next to a freshdirect truck.	Score: 3 Style: short description
Score: 2 Style: short description	 Explanation: Some text is visible on the devices but not the main focus.'	Q: Provide a one-sentence caption for the provided image. A: Five different types of cellular phones next to each other on a table.	Score: 2 Style: short description
Score: 1 Style: detailed description	 Explanation: Minimal relevance as there is text in the image but not utilized.	Q: Considering the presence of multiple bottles in the scene, what issue does this environment illustrate, and how can it affect pedestrians? Answer the question with GPT-T-COCO format. A: The environment in the image shows an issue of littering, as multiple bottles are scattered across the scene near the road. This can negatively affect pedestrians in several ways, such as creating a dirty and unkempt surrounding, posing health risks, ...	Score: 0 Style: detailed description
Score: 0 Style: multi-choice	 Explanation: No text is present in the image.'	Q: You will be given an image of a guitar. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category. Options: (a) Photograph (b) Cartoon (c) Art painting (d) Sketch A: (d) Sketch	Score: 0 Style: multi-choice

Figure 28: Randomly selected samples from the LLaVA-OVSI data pool within each score group for the **optical character recognition** capability, ranging from 5 to 0, as evaluated by GPT-4o, along with corresponding score explanations and SSR predictions.