RankRepair: Decoupling Record Priority and Budget Allocation for training-free multimodal data selection

Anonymous authors
Paper under double-blind review

Abstract

Training vision-language models (VLMs) often requires substantial computational resources and large-scale data, making the selection of compact, high-quality instruction datasets essential for efficient multimodal learning. Selecting such subsets requires balancing individual record priority against the effective allocation of a limited data budget across the overall candidate pool. However, incorporating diversity or global structure into record-level scores does not by itself ensure adequate regional coverage in the selected subset. Regions can remain underrepresented after selection, motivating an explicit adjustment of subset-level allocation. To address this challenge, we introduce RankRepair, a novel two-stage framework that decouples record priority estimation from subset-level allocation adjustment. In the first stage (Rank), complete records are scored by integrating image–instruction alignment with local visual diversity to select an initial high-priority subset. In the second stage (Repair), the framework measures regional shortfalls relative to pool-proportional coverage floors and systematically mitigates them through record swaps while keeping priority evaluations fixed. By imposing explicit bounds on both the number of allowed swaps and the maximum priority loss per swap, RankRepair limits departure from the initial quality-based selection, ensuring each accepted swap strictly reduces the aggregate coverage deficit by one unit. Extensive evaluations on LLaVA-1.5 Mix-665K demonstrate that a 100K training subset selected by RankRepair outperforms the evaluated equal-budget baselines in mean performance across ten multimodal benchmarks. The selected subset retains 96.76% of the full-data mean performance, exceeding the strongest equal-budget baseline by 3.16 percentage points, supporting bounded coverage adjustment as an effective complement to quality-based ranking for visual instruction data selection.

1 Introduction

Visual instruction tuning builds general-purpose multimodal assistants from large and heterogeneous collections of image–conversation records (Liu et al., 2023; 2024a). As these collections grow, training on every available record becomes increasingly expensive, especially when models must be repeatedly fine-tuned across architectures, training recipes, and experimental settings. This has motivated multimodal instruction-data selection, which aims to retain a compact subset of the training pool while preserving useful supervision and downstream performance.

Recent selection methods approach this problem from different perspectives. PRISM (Bi et al., 2025) and ScalSelect (Wu et al., 2026b) exploit representation structure to identify informative training records, while ICONS (Wu et al., 2025) and TIVE (Liu et al., 2025) estimate training value through influence-related signals. DataTailor (Yu et al., 2025) combines informativeness, uniqueness, and representativeness and adapts selection proportions across tasks, while COINCIDE (Lee et al., 2024) allocates budgets across concept–skill clusters using their density and transferability. These methods demonstrate that multimodal data selection depends on both the utility of individual records and the structure of the candidate pool. Under a fixed budget, this can be viewed as two related decisions: which records should receive high priority and how the available training slots should ultimately be distributed across the pool.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These two decisions pose distinct challenges. The first is estimating priority at the level at which visual instruction data provide supervision. A visual instruction corpus is not simply a collection of independent images: the same image can be paired with multiple instructions targeting different objects, visible text, relations, or reasoning skills. Such records share visual content but provide different supervision, so image-level similarity alone is insufficient for determining their relative priority. Conversely, image–instruction correspondence alone does not reveal whether a candidate is visually redundant with nearby records. Record-level priority should therefore reflect both the correspondence between an image and its instruction and the candidate's distinctness within the local visual distribution.

To address these challenges, we propose RankRepair, a two-stage framework that separates record-level priority estimation from post-selection coverage adjustment. The Rank stage addresses the first challenge by combining two complementary signals: image–instruction alignment and local visual diversity. Alignment is measured in a shared space learned with cross-fitted CCA from frozen target-model representations, while local diversity captures visual distinctness from nearby candidates. The two signals are normalized and combined into a record-level priority, from which the initial global top-B subset is selected.

Better record-level ranking, however, does not guarantee a well-allocated selected subset. A global top-B cutoff determines not only which records are preferred but also how the limited budget is distributed across different parts of the candidate pool. The selected mass of a region depends on how its priority distribution interacts with this global cutoff. Regions selected at rates below the overall selection rate lose share relative to the candidate pool, potentially leaving coverage shortfalls even under an informative ranking. Pool proportions provide a coverage reference without additional target data, rather than an assumed optimal training distribution.

The Repair stage addresses this second challenge while preserving the priorities established by Rank. Starting from the global top-B subset, it identifies deficient regions using a fixed visual partition and pool-proportional coverage floors, then performs deficit-directed replacements under explicit limits on both the number of edits and priority loss. RankRepair therefore uses Rank to determine record-level preference and Repair to make bounded post-selection adjustments to the allocation induced by global truncation.

We evaluate RankRepair on LLaVA-1.5 Mix-665K using LLaVA-1.5-7B across ten multimodal benchmarks. With a 100K training budget, RankRepair achieves a mean score of 48.05, outperforming competitive multimodal data-selection baselines while retaining 96.76% of the Full-665K reference performance.

Our contributions are:

- We formulate multimodal instruction-data selection around two related but distinct decisions: estimating priorities for complete image–instruction records and controlling the allocation induced when those priorities are converted into a fixed-budget subset.
- We propose RankRepair, which combines image–instruction alignment and local visual diversity for record-level ranking and then performs bounded, deficit-directed post-selection repair while keeping the priorities fixed.
- We provide controlled evaluations that isolate the contributions of the ranking signals and Repair, including component ablations, matched perturbations, and full-floor controls that distinguish targeted bounded repair from generic subset modification and complete coverage enforcement.

2 Related Work

Multimodal instruction-data selection. Multimodal data selectors use signals of sample utility and dataset structure. PRISM (Bi et al., 2025) identifies anisotropy-induced global semantic drift in visual features and applies implicit re-centering before selection. DataTailor (Yu et al., 2025) combines informativeness, uniqueness, and representativeness, and allocates task-adaptive proportions across task blocks. COINCIDE (Lee et al., 2024) explicitly assigns cluster-level budgets using concept density and transferability before selecting within concept–skill clusters. ScalSelect (Wu et al., 2026b) ranks records by their contribution to a dominant subspace of instruction-conditioned

target-model states. ICONS (Wu et al., 2025) aggregates gradient influence across vision–language tasks, TIVE (Liu et al., 2025) combines instance influence with task difficulty, and DOSE (Wu et al., 2026a) samples according to an off-the-shelf quality–alignment distribution. The candidate unit also varies: PreSel (Safaei et al., 2025) selects images before instruction generation, whereas our setting treats each complete image–instruction pair as a candidate, so records sharing an image can receive different priorities. These methods already account for quality, diversity, or budget allocation. Nevertheless, these criteria address different aspects of selection and are not interchangeable. Visual distinctness alone does not establish whether an image supports its paired instruction, while strong image–instruction agreement can coexist with substantial local redundancy. For example, instructions about different objects in the same image need not have equal visual support, even though an image-only representation treats their visual content alike. Conversely, individually well-aligned records may repeatedly cover a narrow part of the visual pool. Combining correspondence and local diversity therefore helps distinguish useful supervision from repeated visual content without treating all records associated with an image as equivalent. Rank addresses this record-level limitation by combining alignment with local visual diversity. A separate limitation arises when scores are converted into a training subset: diversity-aware ranking does not itself impose regional coverage requirements, and task- or cluster-level quotas can leave imbalances within their groups. Repair addresses this allocation-level issue by identifying residual shortfalls after ranking and correcting them through limited exchanges. The two stages thus connect complementary record-level signals with explicit control over the composition of the selected subset.

Coverage and set-dependent selection. Beyond multimodal instruction tuning, geometric coresets and k-center methods optimize coverage in a chosen representation (Gonzalez, 1985; Sener & Savarese, 2018; Mirzasoleiman et al., 2020). Constrained ranking methods trade ranking utility against representation constraints (Zehlike et al., 2017; Celis et al., 2018), and TSDS (Liu et al., 2024d) aligns a selected subset with a task-specific target distribution. JEST (Evans et al., 2024) shows that joint batch-level scoring can differ from independent example ranking. Coverage objectives alone do not specify how much an existing high-priority subset should change to improve representation. Enforcing coverage without such limits may require replacing many preferred records. Repair makes this tradeoff explicit: it starts from an existing top-B subset and bounds both the number of replacements and their priority loss. This preserves the original ranking as a reference while allowing targeted allocation changes, which we assess against the unrepaired subset and matched perturbations.

3 Method

RankRepair first ranks visual instruction records using image–instruction alignment and local visual diversity, then repairs regional coverage deficits in the global top-B subset. Repair keeps the priorities fixed and bounds both the number of replacements and their per-swap priority loss (Figure 1).

3.1 Problem Setup

Let $\mathcal{U} = \{x_i\}_{i=1}^{N}$ contain image–conversation records. We select B complete records for instruction tuning, where $N \geq 2$ and $1 \leq B \leq N$; records sharing an image remain distinct candidates. Text-only records are fixed across comparisons and appended after selection, outside budget B. Given priorities $q_i \in [0, 1]$, global top-B selection yields

$$S_0 = \operatorname*{arg\,max}_{S \subseteq \mathcal{U},\, |S|=B} \sum_{x_i \in S} q_i. \tag{1}$$

Score ties are resolved deterministically using stable dataset-row identifiers. For a fixed partition $C = \{C_1, \ldots, C_K\}$ of $\mathcal{U}$ into nonempty regions, define pool shares $p_c = |C_c|/N$, regional selection rates $s_c = |S_0 \cap C_c|/|C_c|$, and overall selection rate $\beta = B/N$. The selected share of region C_c is

$$\widehat{p}_c = \frac{|S_0 \cap C_c|}{B} = p_c \frac{s_c}{\beta}. \tag{2}$$

A region gains share relative to the pool when $s_c > \beta$ and loses share when $s_c < \beta$. Maximizing aggregate priority therefore does not explicitly control regional allocation. Repair introduces bounded adjustments to this allocation while preserving the priorities used to construct S_0.

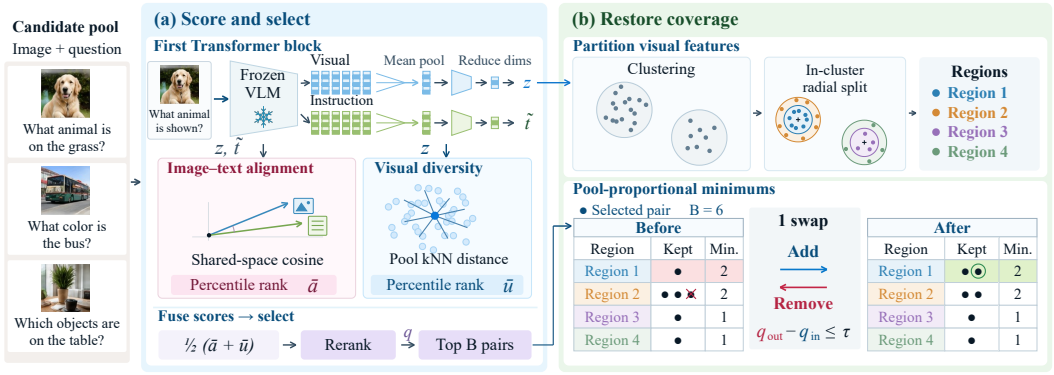

Figure 1: **Rank and Repair.** (a) Rank fuses cross-fitted CCA alignment and visual kNN diversity from frozen Layer 1 features to select global top-B records. (b) Repair uses a fixed partition of the same visual features and pool-proportional floors, keeping B and q fixed. In the swap condition, q_{out} and q_{in} are the removed and added priorities (higher is better), and τ bounds their difference. In the illustrative exchange, $N = 24$, pool counts $(8, 8, 4, 4)$ and $B = 6$ give floors $(2, 2, 1, 1)$. Occupancies change from $(1, 3, 1, 1)$ to $(2, 2, 1, 1)$, reducing the deficit by one while preserving the donor floor. Filled dots denote selected records; the cross and ring mark outgoing and incoming records. Edit and priority-loss limits can leave residual deficits. Images and counts are illustrative, not dataset measurements. Snowflake icon: Freepik/Flaticon.

3.2 Rank: Record Priority Estimation

Rank combines image–instruction correspondence (*Alignment*) with visual distinctness from nearby candidates (*Local Diversity*). For each record x_i, we obtain visual and instruction representations, $v_i \in \mathbb{R}^{d_v}$ and $t_i \in \mathbb{R}^{d_t}$, by separately mean-pooling the visual and text token hidden states at the output of the frozen language model's first Transformer block (Layer 1). We independently apply ℓ_2 normalization and PCA to the two views, yielding compact representations

$$z_i = \phi_v(v_i), \qquad \widetilde{t}_i = \phi_t(t_i), \tag{3}$$

where ϕ_v and ϕ_t include normalization, centering, and PCA projection for the visual and instruction features, respectively. The PCA maps are fitted before CCA cross-fitting and shared across folds.

To measure image–instruction alignment, we use canonical correlation analysis (CCA) to map the independently projected views into a shared space. For cross-fitting, we assign image groups to folds, fit CCA on the remaining folds, and score held-out records. Let $f(i)$ denote the fold containing record i, and let $W_v^{(-f(i))}$ and $W_t^{(-f(i))}$ denote the CCA projections fitted without that fold. We define alignment as the cosine similarity

$$a_i = \cos\left(\left(W_v^{(-f(i))}\right)^\top z_i,\ \left(W_t^{(-f(i))}\right)^\top \widetilde{t}_i\right). \tag{4}$$

A larger a_i indicates stronger image–instruction agreement in the CCA space. Grouping by image ensures that instructions associated with the same image do not appear in both the CCA fitting and held-out sets. We normalize the alignment scores as $\bar{a}_i = r(a_i)$ using empirical percentile ranks. The operator r ranks scores in ascending order over all N candidates, assigns tied scores their average rank, and rescales each one-based rank R to $(R-1)/(N-1)$, yielding $\bar{a}_i \in [0, 1]$.

We complement alignment with local visual diversity, measured in the visual PCA space. Let $\mathcal{N}_k(i)$ contain the indices of the k nearest neighbors of z_i under Euclidean distance, excluding record i, where $1 \le k < N$. The local diversity score is

$$u_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \left\| z_i - z_j \right\|_2. \tag{5}$$

A larger u_i suggests lower local visual redundancy. We independently percentile-normalize diversity over the pool as $\bar{u}_i = r(u_i)$.

With both signals on a percentile scale, we first average their normalized scores with equal weights, then convert the fused scores to final priorities:

$$\widetilde{q}_i = \frac{\bar{a}_i + \bar{u}_i}{2}, \qquad q_i = r(\widetilde{q}_i). \tag{6}$$

The final transformation preserves the fused-score ordering and expresses priority differences on a percentile scale for Repair. Global top-B selection under q_i produces S_0; the priorities then remain fixed.

3.3 Coverage Deficit

We measure coverage using a fixed partition $C = \{C_1, \ldots, C_K\}$ of the candidate pool. To construct it, we reuse the Layer 1 visual representations z_i from the Rank stage and cluster them using MiniBatch k-means. Within each cluster, we sort records by Euclidean distance to the centroid and divide them into approximately equal-sized radial strata. Each stratum defines one region C_c, so every record belongs to exactly one region. The partition is constructed before selection and held fixed across the initial subset, Repair, and the corresponding controls. It is used only to assess and repair allocation and does not enter the priority score q_i.

This partition provides the regions for a pool-proportional coverage reference. Let $N_c = |C_c|$ and $p_c = N_c/N$. We define the coverage floor, or protected occupancy, of region C_c directly from its pool share:

$$\ell_c = \min\{N_c, \lfloor Bp_c \rfloor\}. \tag{7}$$

The floors depend only on the pool shares and budget; regions with $Bp_c < 1$ receive a zero floor. Derived directly from the candidate pool, these floors provide a coverage reference without additional target data. The edit and priority-loss constraints control the extent of adjustment toward this reference.

To quantify how far a selected subset falls below these floors, let $o_c(S) = |S \cap C_c|$ denote the occupancy of region C_c in a size-B subset S. The aggregate coverage deficit is

$$D(S) = \sum_{c=1}^{K} [\ell_c - o_c(S)]_+, \tag{8}$$

where $[x]_+ = \max(x, 0)$. This counts records missing from the regional floors; regions at or above their floors contribute zero.

3.4 Repair: Bounded Allocation Adjustment

Starting from S_0, Repair replaces selected records from regions above their floors with unselected records from deficient regions. Let S_t denote the subset after t accepted swaps. Each swap

$$S_t = (S_{t-1} \setminus \{x_{o_t}\}) \cup \{x_{i_t}\} \tag{9}$$

preserves $|S_t| = B$, where $x_{o_t} \in S_{t-1}$ and $x_{i_t} \notin S_{t-1}$ are the outgoing and incoming records.

We limit the change from the initial selection through two constraints. An edit fraction $\epsilon \in [0, 1]$ permits at most $E = \lfloor \epsilon B \rfloor$ swaps. Each swap must also satisfy

$$\Delta q_t = q_{o_t} - q_{i_t} \leq \tau, \tag{10}$$

where $\tau \geq 0$ bounds the percentile-priority loss per replacement.

At each step, we choose the unblocked region with the largest positive deficit $[\ell_c - o_c(S_{t-1})]_+$ and take its highest-priority available unselected record. The donor is the lowest-priority eligible selected record, whose region C_a must satisfy $o_a(S_{t-1}) > \ell_a$. Region ties are resolved by region index; record ties use stable dataset-row identifiers. Inserted records cannot be removed, and removed records cannot be reinserted. If the pair satisfies Equation 10, we accept it, update occupancies, and repeat using the new deficits. Otherwise, or if no candidate pair exists, we block that region and continue with the others. Repair stops when all floors are met, E swaps have been accepted, or all remaining deficient regions are blocked.

The resulting changes can be bounded in terms of the accepted swaps. For the final subset S, define the replacement count $m(S) = |S \setminus S_0|$ and the average priority loss per selected record

$$L_q(S) = \frac{1}{B}\left(\sum_{x_i \in S_0} q_i - \sum_{x_j \in S} q_j\right). \tag{11}$$

Each accepted swap preserves the donor region's floor and reduces the recipient region's deficit by one. Since no record is edited twice, T accepted swaps yield

$$\begin{aligned} D(S_T) = D(S_0) - T, \qquad m(S_T) = T \le E, \\ 0 \le L_q(S_T) \le \frac{T\tau}{B} \le \epsilon\tau. \end{aligned} \tag{12}$$

The loss is nonnegative because S_0 maximizes total priority; its upper bound follows by summing per-swap losses. Thus, T swaps correct T units of coverage deficit while preserving at least $B - E$ original records and limiting average priority loss to $\epsilon\tau$. These bounds apply to accepted swaps. Appendix C.6 compares greedy scheduling with an exact constrained allocation oracle.

As a control for the bounded procedure, we also enforce every floor without repair limits, selecting the highest-priority feasible size-B subset (Appendix B). This full-floor control allows the experiments to compare partial coverage repair with complete enforcement of the same coverage reference.

4 Experiments

4.1 Experimental Setup

Our main experiments evaluate RankRepair on LLaVA-1.5 Mix-665K across selection budgets, with controlled ablations examining how ranking and targeted coverage repair contribute to downstream performance. Each 100K training subset contains 93,884 visual records from this pool and the same 6,116 text-only records. We train LLaVA-1.5-7B for one epoch using the same initialization, optimizer, prompt format, and evaluation configuration across methods. At larger budgets, all methods retain the same text-only subset, so the visual selection budget B is the total training budget minus 6,116. With one training epoch throughout, larger budgets entail more computation; within-budget protocols remain fixed. RankRepair uses frozen Layer 1 visual and instruction representations. The coverage partition reuses the same visual representations used for local diversity. For RankRepair and its controls, features and the partition are held fixed across budgets; the pool-proportional floors ℓ_c and edit cap $E = \lfloor \epsilon B \rfloor$ are recomputed for each B. Unless stated otherwise, Repair uses $\epsilon = 0.02$ and $\tau = 0.05$. Rank uses 32-dimensional PCA, five-fold cross-fitted CCA with 16 components and regularization 0.05, and $k = 20$ neighbors. Coverage uses 128 clusters and three radial strata. Appendix A provides implementation details.

Evaluation covers MMBench-EN/CN (Liu et al., 2024b), MME (Fu et al., 2025), POPE (Li et al., 2023), MM-Vet (Yu et al., 2024), MMMU (Yue et al., 2024), OCRBench (Liu et al., 2024c), AI2D (Kembhavi et al., 2016), MMStar (Chen et al., 2024), and RealWorldQA (xAI, 2024). Let s_d denote the normalized 0–100 score on task d. For OCRBench, we divide the raw 0–1000 score by 10. For MME, we use

$$s_{\mathrm{MME}} = \frac{1}{2}\left(100\,\frac{\mathrm{MME_P}}{2000} + 100\,\frac{\mathrm{MME_C}}{800}\right).$$

Our primary aggregate metric is the unweighted ten-task mean, $M = \frac{1}{10}\sum_{d=1}^{10} s_d$.

Random samples visual records uniformly without replacement. Source-prop. fixes pool-proportional source quotas before sampling within each source. We compare against PRISM (Bi et al., 2025), DataTailor (Yu et al., 2025), ScalSelect (Wu et al., 2026b), ICONS (Wu et al., 2025), COINCIDE (Lee et al., 2024), and TIVE (Liu et al., 2025), together with length and perplexity as inexpensive scalar-ranking controls. Rank-only uses our Rank priority without repair, while RankRepair applies Repair to the same ranking. Alignment-only and local-diversity-only variants test the two scoring components separately. A boundary perturbation matches Repair's replacement count and aggregate priority loss by adjusting record sets around the Rank cutoff, without regional coverage information (Appendix B). Full-floor enforcement satisfies every regional floor without repair limits; we evaluate its selection structure and downstream performance. Full-665K is a full-data reference, not an equal-budget selector. The appendix gives training details.

4.2 Main Results

Table 1: **100K comparison across ten benchmarks.** Task scores are means across three runs; Mean reports the ten-task average ± sample standard deviation across seeds. Retention is the displayed Mean divided by Full-665K's Mean. Bold/underline mark the best/second-best subset results.

Method	MMB-E	MMB-C	MME	POPE	MM-Vet	MMMU	OCR	AI2D	MMStar	RWQA	Mean	Retention
LLaVA-1.5 (Full-665K)	63.57	44.85	59.88	86.45	37.16	33.89	30.80	51.94	35.07	52.94	49.66	100.00%
Perplexity	31.74	32.47	52.51	84.81	30.95	31.64	18.60	38.02	29.38	42.75	39.29 ± 0.28	79.12%
Length	49.10	36.43	50.33	81.82	29.76	30.92	13.60	37.44	28.91	42.27	40.06 ± 0.31	80.67%
Random	52.34	35.21	51.06	80.70	30.84	31.75	17.90	50.78	29.47	43.02	42.31 ± 0.24	85.20%
ScalSelect	56.53	42.10	53.41	84.26	34.77	33.56	26.30	48.06	30.80	43.65	45.34 ± 0.23	91.30%
Source-prop.	57.90	30.15	50.46	**84.86**	**38.99**	34.11	28.80	50.84	**33.60**	49.07	45.88 ± 0.27	92.39%
TIVE	56.92	43.81	53.74	83.65	33.21	33.36	26.40	49.85	31.72	46.68	45.93 ± 0.21	92.49%
ICONS	56.84	43.72	53.67	84.21	33.58	33.47	26.80	49.72	31.84	46.91	46.08 ± 0.19	92.79%
COINCIDE	56.84	44.21	53.76	84.08	33.71	33.35	27.10	49.86	31.82	47.01	46.17 ± 0.18	92.97%
DataTailor	58.08	48.71	**56.26**	82.70	31.80	33.20	27.90	51.36	31.40	40.40	46.18 ± 0.20	92.99%
PRISM	57.47	45.16	51.10	84.85	34.11	**36.22**	**29.00**	49.87	32.53	44.53	46.48 ± 0.15	93.60%
RankRepair	**59.71**	**49.66**	54.39	84.29	34.50	35.56	28.40	**51.55**	31.93	**50.50**	**48.05 ± 0.22**	**96.76%**

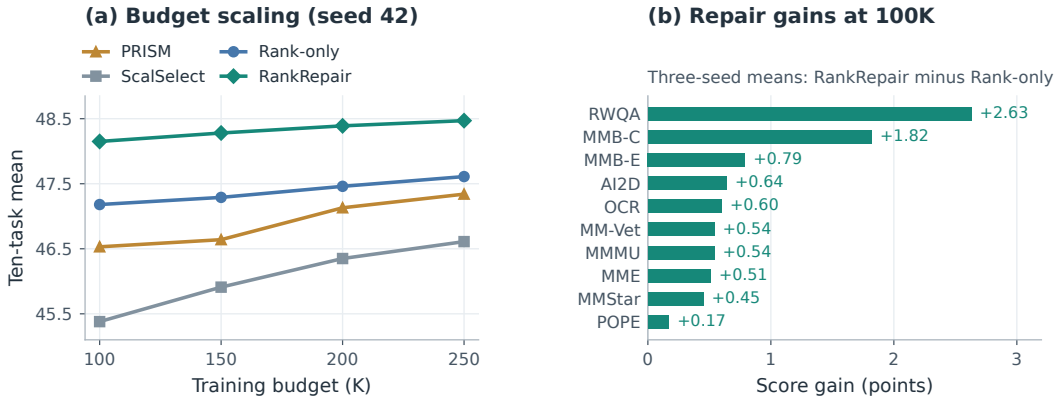

Figure 2: **Budget scaling and Repair gains.** (a) Single-run ten-task Mean across budgets (seed 42). (b) Three-seed mean task-score changes from Rank-only to RankRepair at 100K. Exact scores are in Appendix Tables 6 and 7.

At 100K, RankRepair achieves the highest selected-subset ten-task Mean of 48.05 ± 0.22, retaining 96.76% of the Full-665K reference (49.66; Table 1). Its three-seed mean exceeds PRISM, COINCIDE, and ScalSelect by 1.57, 1.88, and 2.71 points, respectively. The retention advantage over the strongest competing subset, PRISM, is 3.16 percentage points (96.76% versus 93.60%). RankRepair leads the subset methods on MMBench-EN, MMBench-CN, AI2D, and RealWorldQA, while other selectors remain stronger on the remaining benchmarks. For example, Source-prop. performs best on POPE, MM-Vet, and MMStar, and PRISM leads on MMMU and OCRBench. The higher aggregate reflects a balance across tasks, not dominance on every benchmark. Full-665K retains the higher Mean, while RankRepair preserves most of its performance with fewer training records.

In the seed-42 budget runs, RankRepair rises from 48.15 at 100K to 48.47 at 250K and Rank-only from 47.18 to 47.61, with paired gains of +0.97, +0.99, +0.93, and +0.86 (Figure 2(a)). Within-budget protocols are fixed; cross-budget trends combine sample count and one-epoch training computation. Figure 2(b) shows positive changes on all ten tasks for the three-seed 100K comparison, with the largest gains on RealWorldQA (+2.63) and MMBench-CN (+1.82), versus +0.17 on POPE. The gains vary across tasks; this comparison isolates allocation changes under a fixed ranking, complementing comparisons against external selectors.

Table 2: **Ranking and allocation controls at 100K.** Downstream scores are single-run results (seed 42). D is residual coverage deficit; Loss is $10^4 L_q$ relative to Rank-only. Δ is relative to Rank-only. Dashes mark inapplicable or unreported quantities. Matching is within numerical tolerance; losses are rounded to two decimals.

Selection	Residual D	Swaps	Loss	Mean	Δ
Alignment-only top-B	6,104	–	–	46.31	-0.87
Local-diversity top-B	6,287	–	–	45.96	-1.22
Rank-only	5,812	0	0.00	47.18	0.00
Matched perturbation	5,706	1,877	4.39	46.91	-0.27
Source-matched control	–	1,877	4.39	47.03	-0.15
RankRepair	3,935	1,877	4.39	**48.15**	**+0.97**
Full-floor enforcement	0	5,812	30.11	47.74	+0.56

4.3 Ablation Study

Table 2 compares selection structure and performance. Alignment-only and local-diversity-only use their own top-B rankings; other controls fix the fused Rank priority and change only allocation.

Rank-only improves over alignment alone by 0.87 points and over local diversity alone by 1.22. Its initial coverage deficit is also smaller (5,812 versus 6,104 and 6,287), but remains nonzero across 74 of 384 regions with positive floors. The complementary signals yield the strongest of these three downstream results without ensuring every regional floor. We therefore test allocation changes with the fused priority fixed to distinguish their contribution from that of scoring.

Repair reduces the deficit by 1,877 records (32.3%), retaining 98.00% of the original visual subset. The boundary perturbation matches its 1,877 replacements and aggregate priority loss, but removes only 106 records of deficit (1.8%). Its Mean falls to 46.91, whereas Repair reaches 48.15. The source-matched control additionally matches per-source inserted/removed counts up to discrete feasibility and reaches 47.03, still 1.12 points below Repair. Its maximum source-count error is three records and its loss-matching error is 6.2×10^{-8} (Appendix C.4). These comparisons support targeted repair beyond matched boundary or coarse source-composition changes, without identifying individual records' contributions or region-specific task benefits. Appendix C.6 compares the greedy scheduler with an exact constrained allocation oracle, and Appendix C.7 audits the dataset-source composition of the 1,877 accepted exchanges.

Full-floor enforcement eliminates the deficit, but uses 3.1× as many replacements and 6.9× the priority loss of Repair. Its Mean is 47.74, below Repair by 0.41 points. Thus, minimizing the measured coverage deficit alone does not maximize downstream performance in this comparison. Repair mainly reduces shortfall severity: the maximum regional deficit falls from 261 to 192, while deficient regions fall from 74 to 71. Appendix Tables 5 and 7 provide overlap and three-seed per-task results, respectively. The latter confirm a mean gain of 0.87 points (48.05 versus 47.18). Most deficient regions thus remain below their floors: the observed gain accompanies reduced shortfall severity, not complete coverage.

To relate correction strength to downstream benefit, we fix Rank scores and the partition, varying ϵ at $\tau = 0.05$ and τ at $\epsilon = 0.02$.

Increasing the edit cap from 1% to 2% to 5% reduces the deficit by 16.1%, 32.3%, and 33.1%, respectively, with scaled priority losses $10^4 L_q$ of 2.08, 4.39, and 4.56. Their downstream Means are 47.93, 48.15, and 48.22, respectively. The 5% cap admits only 1,926 actual swaps (2.05% of the visual subset), yielding 0.07 points over the default. Full-floor enforcement instead falls below these bounded settings despite eliminating the measured deficit. At fixed $\epsilon = 0.02$, tightening τ from 0.05 to 0.02 reduces accepted swaps from 1,877 to 1,672, while the Mean changes only slightly from 48.15 to 48.19. At $\tau = 0.10$, the reported run also reaches the 1,877-edit cap, with Mean 48.17. Nearby settings yield similar observed Means; the differences are small relative to the main Repair gain and do not establish a reliable ordering among the nearby settings. The sweeps show how either constraint can limit correction: the loss bound can prevent full use of the edit cap, while a reached cap prevents further correction even under a looser loss bound.

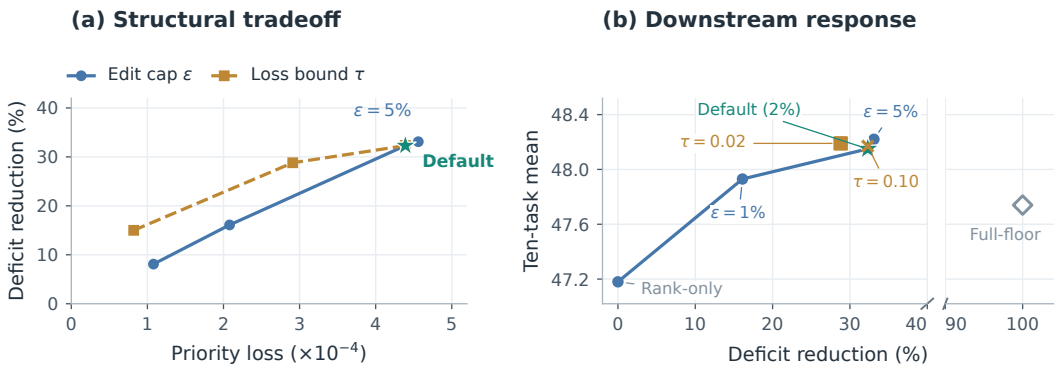

Figure 3: **Coverage correction and downstream performance.** (a) One-factor edit-cap and loss-bound sweeps; the star marks the default. (b) Observed downstream scores; the line connects edit-cap settings, with loss-bound and full-floor controls plotted separately. Exact values: Appendix Tables 13–14 and Table 2.

Table 3: **Second model and training pool with Qwen3-VL-4B on Vision-Flan 57K.** Results are reported separately from the main LLaVA comparison. OCRBench scores are divided by 10; all displayed scores use the 0–100 scale.

Method	MMB-E	MMB-C	POPE	MMVet	MMMU	OCR	AI2D	MMStar	RWQA
PRISM	78.62	78.94	85.31	44.18	47.86	67.20	78.93	54.47	61.42
DataTailor	78.31	78.57	85.08	43.86	47.42	66.60	78.61	54.11	61.03
Rank-only	79.82	80.31	86.25	45.91	49.15	69.10	80.25	55.88	63.02
RankRepair	**80.07**	**80.67**	**86.47**	**46.38**	**49.44**	**69.40**	**80.57**	**56.27**	**63.53**

4.4 Additional Model, Transfer, and Efficiency

Table 3 evaluates Qwen3-VL-4B (Bai et al., 2025) with 57K records selected from a 186,103-record Vision-Flan pool (Xu et al., 2024). Rank-only exceeds PRISM and DataTailor on all nine displayed tasks, and Repair improves further on each. Both model and data change together. OCRBench is divided by 10; MME is omitted, so we do not report the main ten-task Mean (Appendix F).

A 64-cluster, three-stratum partition reaches 47.96 versus the default 128-cluster setting's 48.15, with 96.7% subset overlap. Other partition settings and neighborhood-size results are in Appendix D. Repair also reduces selection-side deficit under PRISM, ScalSelect, DataTailor, and Rank (5.6%, 10.5%, 31.2%, and 32.3%). A downstream transfer check further applies the same fixed Repair operator to PRISM and DataTailor priorities; both improve over their unrepaired subsets while remaining below RankRepair (Appendix Table 10).

After shared 3.0-hour Layer 1 feature preparation, cached selection takes 2.1 minutes for Rank-only and 3.5 for RankRepair: Repair adds 1.4 minutes but no training or inference operations. Appendix E distinguishes end-to-end timings from operations on cached final scores.

5 Conclusion

RankRepair combines alignment–diversity ranking with bounded coverage repair while keeping priorities fixed. On LLaVA-1.5 Mix-665K, its 100K subset achieves a three-seed ten-task mean of 48.05 ± 0.22. In the three-seed allocation ablation, Repair improves the mean by 0.87 points and reduces the coverage deficit by 32.3% while retaining 98% of the initial visual subset. Gains exceed the matched perturbation and remain positive across evaluated budgets, supporting bounded repair as a complement to ranking. The edit and priority-loss constraints provide direct control over coverage adjustment, enabling improvements through limited changes to a high-priority training subset. The coverage reference inherits the candidate pool's composition, and the guarantees concern allocation and priority loss rather than downstream accuracy.

AI Use Statement

Generative AI tools assisted with manuscript editing, discussion of reported experimental results, and the review and exposition of mathematical arguments. The authors are responsible for the final text, claims, and figures.

Reproducibility Statement

The method and its assumptions are described in Section 3, and the experimental setup and evaluation protocol are described in Section 4. The appendix provides further details on the coverage-floor and repair procedure, training configuration, and additional experimental analyses.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. PRISM: self-pruning intrinsic selection method for training-free multimodal data selection. *CoRR*, abs/2502.12119, 2025. doi: 10.48550/ARXIV.2502.12119. URL https://doi.org/10.48550/arXiv.2502.12119.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella (eds.), *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, Prague, Czech Republic, July 9-13, 2018*, volume 107 of *LIPIcs*, pp. 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPICS.ICALP.2018.28. URL https://doi.org/10.4230/LIPIcs.ICALP.2018.28.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ff8d608f6dcebec401df78ca76617e95-Abstract-Datasets_and_Benchmarks_Track.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz,

Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5. URL https://doi.org/10.1016/0304-3975(85)90224-5.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pp. 235–251. Springer, 2016. doi: 10.1007/978-3-319-46493-0_15. URL https://doi.org/10.1007/978-3-319-46493-0_15.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5060–5080. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.291. URL https://doi.org/10.18653/v1/2024.emnlp-main.291.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

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. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024b. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of OCR in large multimodal models. *Sci. China Inf. Sci.*, 67(12), 2024c. doi: 10.1007/S11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: data selection for task-specific model finetuning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10*

- *15, 2024*, 2024d. URL http://papers.nips.cc/paper_files/paper/2024/hash/13848b5893119ff772b69812c95914fa-Abstract-Conference.html.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: High-value data selection for visual instruction tuning. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau (eds.), *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 3712–3721. ACM, 2025. doi: 10.1145/3746027.3755160. URL https://doi.org/10.1145/3746027.3755160.

Baharan Mirzasoleiman, Jeff A. Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960. PMLR, 2020. URL http://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14247–14256. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01329. URL https://openaccess.thecvf.com/content/CVPR2025/html/Safaei_Filter_Images_First_Generate_Instructions_Later_Pre-Instruction_Data_Selection_for_CVPR_2025_paper.html.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Biao Wu, Yiwu Zhong, Meng Fang, and Ling Chen. DOSE: data selection for multi-modal llms via off-the-shelf models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 897–911. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.FINDINGS-ACL.45. URL https://doi.org/10.18653/v1/2026.findings-acl.45.

Changti Wu, Jiahuai Mao, Yuzhuo Miao, Shijie Lian, Bin Yu, Xiaopeng Lin, Cong Huang, Lei Zhang, and Kai Chen. Scalselect: Scalable training-free multimodal data selection for efficient visual instruction tuning. *CoRR*, abs/2602.11636, 2026b. doi: 10.48550/ARXIV.2602.11636. URL https://doi.org/10.48550/arXiv.2602.11636.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. ICONS: influence consensus for vision-language data selection. *CoRR*, abs/2501.00654, 2025. doi: 10.48550/ARXIV.2501.00654. URL https://doi.org/10.48550/arXiv.2501.00654.

xAI. RealWorldQA. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-Flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15271–15342. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.905. URL https://aclanthology.org/2024.findings-acl.905/.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 155–165. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00022. URL https://doi.org/10.1109/ICCV51701.2025.00022.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (eds.), *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pp. 1569–1578. ACM, 2017. doi: 10.1145/3132847.3132938. URL https://doi.org/10.1145/3132847.3132938.

A Rank and Partition Details

Rank representations. Rank uses the mean-pooled visual-token and instruction-token hidden states from Transformer block 1 of the frozen target model. Block 1 denotes `hidden_states[1]`, not the input embedding. The two views are independently ℓ_2-normalized and projected to 32 dimensions with randomized PCA fitted on 30,000 records.

Cross-fitted alignment. Image identifiers define five deterministic folds. For each held-out fold, regularized CCA is fitted on the other four folds and truncated to 16 canonical components, with covariance regularization 0.05. Held-out cosine agreement is converted to a tie-aware percentile $\bar{a}_i = r(a_i)$.

Local diversity and fusion. The default Local-Diversity neighborhood is $k = 20$:

$$u_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \|z_i - z_j\|_2, \qquad \bar{u}_i = r(u_i).$$

The final priority is

$$\widetilde{q}_i = \frac{\bar{a}_i + \bar{u}_i}{2}, \qquad q_i = r(\widetilde{q}_i).$$

Coverage partition. The coverage partition is constructed without using the Rank scores. It reuses the same Layer 1 visual representations z_i used for local diversity, clustered with MiniBatch k-means using 128 clusters, batch size 4096, one initialization, and at most 100 iterations. Within each cluster, records are ordered by centroid distance and split into three approximately equal radial strata, producing 384 mutually exclusive regions. The partition is fixed before selection and shared by Rank-only, RankRepair, matched perturbation, and full-floor enforcement.

B Coverage Floors and Repair

For region C_c, let $N_c = |C_c|$ and $p_c = N_c/N$. The coverage floor is

$$\ell_c = \min\{N_c, \lfloor Bp_c \rfloor\}.$$

No floor-strength or minimum-support parameter is used; regions with $Bp_c < 1$ automatically receive zero floor.

Starting from $S_0 = \mathrm{Top}_B(q)$, Repair chooses the unblocked deficient region with the largest current deficit, inserts its highest-priority available unselected record, and removes the lowest-priority eligible selected record whose donor region remains at or above its floor. A swap is accepted only if

$$q_{\text{out}} - q_{\text{in}} \le \tau,$$

and at most $E = \lfloor \epsilon B \rfloor$ swaps are accepted. The default is $(\epsilon, \tau) = (0.02, 0.05)$. Inserted records are not removed again and removed records are not reinserted. Every accepted swap reduces the aggregate deficit by exactly one, hence

$$D(S_T) = D(S_0) - T, \qquad L_q(S_T) \le \epsilon\tau.$$

Matched boundary perturbation. The matched perturbation starts from the same Rank-only subset S_0 and constructs a control subset using only Rank priority around the global cutoff. Let $O \subseteq S_0$ be the records removed from the selected subset and $I \subseteq \mathcal{U} \setminus S_0$ the records inserted into it. We require

$$|O| = |I| = T,$$

where T equals the number of replacements made by RankRepair. For any fixed O and I, the aggregate priority loss is

$$L_q(O, I) = \frac{1}{B} \left(\sum_{i \in O} q_i - \sum_{j \in I} q_j \right).$$

Loss matching changes the membership of O and/or I, since re-pairing fixed records leaves L_q unchanged. We construct the control over boundary *record sets*, not boundary pairs. The procedure is

deterministic for a fixed seed. We sort the selected side by increasing distance from the cutoff and the unselected side by the same criterion, initialize a boundary pool of $4T$ candidates on each side, and solve the cardinality-constrained loss-matching problem by a seeded random-restart local search. Each restart alternates updates to O and I, replacing one member with an unused candidate from its respective boundary pool whenever this reduces

$$\delta_L = \left| L_q(O, I) - L_q^{\text{Repair}} \right|. \tag{13}$$

We use seed 42, at most 2,000 local-search iterations per restart, and an absolute tolerance of 10^{-7}. If the tolerance is not reached after 20 restarts, the candidate pool is expanded by T records on each side and the search is repeated. Expansion continues until the tolerance is reached or all boundary candidates are exhausted; in the latter case, we retain the lowest-error feasible solution and report its achieved δ_L rather than calling it an exact match. At 100K, both aggregate losses round to 4.39×10^{-4}, with $\delta_L = 4.0\times10^{-8}$. The control matches aggregate loss only; it does not impose Repair's per-swap constraint on individual replacements. The construction never uses region identities, coverage floors, regional occupancies, or deficit values.

Full-floor enforcement. The floors are feasible because $\ell_c \le N_c$ and $\sum_c \ell_c \le B$. Take the top ℓ_c records from each region and fill the remaining slots by global priority. An optimal feasible subset can be chosen to include these regional prefixes: replacing any lower-ranked record in the same region preserves feasibility and cannot decrease total priority. The global fill then maximizes the remaining priority. Under the same total ordering as S_0, the mandatory prefixes add exactly $D(S_0)$ records from outside S_0; all remaining fill records come from S_0. Thus this construction uses exactly $D(S_0)$ replacements, the minimum needed to satisfy the floors.

Same-image diagnostics. The archived diagnostic includes 87,738 image groups with multiple records and multiple distinct normalized instruction strings. For an eligible group G, instruction heterogeneity is the mean pairwise cosine distance between the archived normalized text representations:

$$H_G = \frac{2}{|G|(|G|-1)} \sum_{\substack{i<j \\ i,j\in G}} \left(1 - \cos(\widetilde{t}_i, \widetilde{t}_j)\right).$$

Let s_i^{raw} denote the raw priority score stored in that diagnostic, before the final percentile transformation. We use the population variance within each group:

$$\bar{s}_G = \frac{1}{|G|} \sum_{i\in G} s_i^{\text{raw}}, \qquad V_G = \frac{1}{|G|} \sum_{i\in G} (s_i^{\text{raw}} - \bar{s}_G)^2.$$

Across the eligible groups, the Pearson correlation between H_G and V_G is 0.314. A group has low score variance when $V_G < 10^{-6}$. There are 3,941 such groups out of 87,738, giving 4.49%, with each group weighted equally. Both the correlation and the low-variance fraction are computed on archived raw scores before percentile normalization. The archived top-B subset partially selects 34,312 image groups.

C Additional Experimental Results

C.1 Training Configuration

Unless otherwise stated, all 100K subsets are used to fine-tune LLaVA-1.5-7B for one epoch under the same training recipe. Table 4 summarizes the fixed configuration used for the reported comparisons. For Table 1, the three training seeds are 42, 123, and 3407; we report their mean and sample standard deviation. Table 7 reports three-seed task averages. Table 2 and Figure 3 retain single-run results with seed 42.

Table 4: **Fixed downstream training configuration.** Selection methods change only the training subset; the optimization recipe is held fixed.

Item	Configuration
Base language model	Vicuna-7B-v1.5 / LLaVA-1.5-7B
Vision tower	CLIP ViT-L/14-336
Training epochs	1
Fine-tuning	LoRA
LoRA rank / alpha	128 / 256
Distributed optimizer	DeepSpeed ZeRO-3
Training GPUs	8 GPUs
Random seed	42 unless otherwise stated
Selection budget	100K unless otherwise stated

C.2 Coverage and Budget Results

Tables 5 and 6 provide the exact coverage values summarized in Table 2 and the budget curves shown in Figure 2.

Table 5: **Coverage repair at 100K.** Residual D is defined in Equation 8; Loss reports $10^4 L_q(S)$ from Equation 11. Reduction is $100[D(S_0) - D(S)]/D(S_0)$, and Retained is $100|S \cap S_0|/B$. All rows use the same priorities, partition, and floors. The matched perturbation matches Repair in replacement count and aggregate L_q by adjusting the outgoing and incoming boundary record sets; it does not use coverage information.

Selection	Swaps	Residual D	Reduction (%)	Loss	Retained (%)
Rank-only	0	5,812	0.0	0.00	100.00
Matched perturbation	1,877	5,706	1.8	4.39	98.00
RankRepair	1,877	3,935	32.3	4.39	98.00
Full-floor enforcement	5,812	0	100.0	30.11	93.81

Table 6: **Ten-task Mean across selection budgets.** Each cell reports one method–budget run with training seed 42; Table 1 instead averages three training seeds.

Method	100K	150K	200K	250K
PRISM	46.53	46.64	47.13	47.34
ScalSelect	45.38	45.91	46.35	46.61
Rank-only	47.18	47.29	47.46	47.61
RankRepair	48.15	48.28	48.39	48.47

C.3 Allocation-Control Downstream Results

Table 7 expands the main allocation ablation to the individual benchmarks, averaging three training seeds rather than the single run in Table 2. All four rows use the same Rank priority and 100K training protocol; only the subset-allocation rule changes.

Table 7: **Downstream results for allocation controls at 100K.** Task scores and Mean are averages across three training seeds.

Selection	MMB-E	MMB-C	MME	POPE	MM-Vet	MMMU	OCR	AI2D	MMStar	RWQA	Mean
Rank-only	58.92	47.84	53.88	84.12	33.96	35.02	27.80	50.91	31.48	47.87	47.18
Matched perturbation	58.41	47.26	53.61	83.94	33.71	34.86	27.50	50.63	31.19	47.99	46.91
RankRepair	59.71	49.66	54.39	84.29	34.50	35.56	28.40	51.55	31.93	50.50	48.05
Full-floor enforcement	59.08	48.92	54.10	84.06	34.21	35.18	28.10	51.22	31.71	50.82	47.74

Full-floor enforcement improves over Rank-only but remains below RankRepair, while changing 6.19% of the visual subset versus 2.00% for RankRepair. RankRepair achieves a higher downstream Mean with fewer replacements and lower priority loss in this comparison.

C.4 Subset-Composition Audit

We audit composition within the selected $B = 93{,}884$ visual records, relative to Rank-only; the 6,116 text-only records are fixed across methods. Source shift is the maximum absolute percentage-point change of any original data source relative to Rank-only; task shift is defined analogously over the coarse task-family tags. Text length is measured in tokenizer tokens. Shared-image records count selected records whose image identifier occurs more than once in the selected subset.

Table 8: **Composition audit for 100K allocation controls.**

Selection	Source shift	Task shift	Mean tokens	Unique images	Shared-image records
Rank-only	0.00 pp	0.00 pp	92.8	71,436	31,284
Matched perturbation	0.28 pp	0.34 pp	93.1	71,502	31,196
RankRepair	0.31 pp	0.41 pp	93.4	71,618	31,071
Full-floor	0.86 pp	0.93 pp	94.2	71,901	30,644

Because RankRepair changes source composition slightly more than the basic matched perturbation, we additionally construct a source-composition-matched control without region or deficit information. It matches the Repair swap count, aggregate priority loss, and per-source inserted/removed counts up to the discrete feasibility of the boundary pool. The achieved maximum source-count error is 3 records, $\delta_L = 6.2 \times 10^{-8}$, and its ten-task Mean is 47.03. This remains below RankRepair's 48.15, a 1.12-point difference after matching these coarse source-composition statistics.

C.5 Transfer Across Priority Rankings

Repair is defined as a post-ranking repair operator and does not require the underlying priority to be produced by Rank. We therefore apply the same coverage partition and default Repair configuration $(\epsilon, \tau) = (0.02, 0.05)$ to three external priority rankings. Repair starts from each method's original selected subset, and its percentile-normalized priority is used to evaluate the swap-loss constraint. DataTailor retains its task-adaptive allocation. For such quota-based initial subsets, the deficit-reduction and loss upper bounds still apply; the nonnegative-loss guarantee in Equation 12 requires a global top-B initial subset.

Table 9: **Selection-only transfer of Repair at 100K.** Before and After denote aggregate coverage deficit before and after repair. Reduction is relative to the corresponding initial deficit. Loss reports $10^4 L_q(S)$, and Retained is overlap with the method's original selected visual subset.

Priority	Before D	After D	Reduction	Swaps	Loss	Retained
PRISM	6,486	6,123	5.6%	363	0.84	99.61%
ScalSelect	7,118	6,371	10.5%	747	1.71	99.20%
DataTailor	6,024	4,147	31.2%	1,877	4.52	98.00%
Rank	5,812	3,935	32.3%	1,877	4.39	98.00%

Repair reduces the coverage deficit under all four tested priority rankings, although the amount of feasible correction varies substantially with the initial ranking. ScalSelect starts with the largest deficit (7,118), while PRISM admits the smallest relative correction under the default edit and loss constraints, reducing its deficit by 5.6%. ScalSelect reaches a 10.5% reduction, whereas DataTailor and Rank permit substantially larger reductions of 31.2% and 32.3%, respectively. These results show that the same repair operator can produce feasible structural corrections under different initial rankings, while the attainable correction depends on both the initial allocation and the priority structure near the selection boundary. These selection-side results establish that the same constrained operator is feasible under different priorities; Table 10 adds a downstream check for two representative external rankings.

Table 10: **Downstream transfer of Repair across external priorities at 100K.** The Repair configuration and coverage partition are unchanged. Mean is the ten-task aggregate under the same LLaVA-1.5-7B training protocol.

Priority	Original Mean	+Repair Mean	Δ
PRISM	46.53	46.91	+0.38
DataTailor	46.22	46.76	+0.54
Rank	47.18	**48.15**	+0.97

The downstream check follows the same ordering as the selection-side result: Repair improves both tested external priorities without changing their scoring functions. The gain is smaller for PRISM, for which only 363 feasible swaps are admitted by the fixed constraints, and larger for DataTailor, for which the edit cap is reached. These results support transfer of the post-ranking operator beyond the proposed Rank score, while RankRepair remains the strongest configuration in this comparison.

C.6 Exact Constrained Repair Oracle

The guarantees in Equation 12 apply to accepted swaps and do not imply global optimality of the greedy schedule. To quantify the scheduling gap without retraining a VLM, we construct an exact min-cost-flow oracle under the same regional floors, edit cap E, and per-swap admissibility condition $q_{\text{out}} - q_{\text{in}} \leq \tau$. Donor regions provide capacity equal to their occupancy surplus above the floor, deficient regions receive capacity equal to their current deficit, and feasible donor–recipient record pairs form unit-capacity edges. Each outgoing and incoming record also has unit capacity, ensuring that no record participates in two swaps. We first maximize flow up to E and then, conditional on that maximum number of repairs, minimize total priority loss. This lexicographic objective directly tests whether the greedy schedule leaves feasible deficit reduction unused and how much priority loss is caused by its ordering.

Table 11: **Greedy Repair versus an exact constrained allocation oracle.** Both use the 100K Rank-only subset and the same $(\epsilon, \tau) = (0.02, 0.05)$ constraints.

Scheduler	Swaps	Residual D	Loss ($\times 10^{-4}$)	Runtime
Greedy RankRepair	1,877	3,935	4.39	1.34 s
Exact min-cost flow	1,877	3,935	4.21	38.6 s

The exact oracle finds no additional feasible repair under the fixed edit and per-swap constraints: both schedules use all 1,877 allowed edits and reduce the deficit by the same 32.3%. Its lower aggregate priority loss (4.21 versus 4.39×10^{-4}) indicates a small scheduling cost for the greedy rule, while the attainable deficit reduction is unchanged in this instance. We use the greedy schedule in the main method because its role is bounded targeted repair rather than solving a global allocation objective exactly.

C.7 Source Composition of Accepted Exchanges

We compare the dataset sources of the 1,877 records removed from Rank-only with those of the 1,877 records inserted by Repair. Source labels are used for this post-selection analysis and do not enter the Repair decisions.

Table 12: **Dataset-source composition of the 1,877 accepted exchanges.** Removed and Inserted report source shares within their respective sets of 1,877 records. Δ is Inserted minus Removed in percentage points (pp), not the change in share of the full selected subset.

Source	Removed (%)	Inserted (%)	Δ (pp)
COCO	48.7	33.2	-15.5
GQA	14.8	22.0	+7.2
OCR-VQA	13.6	11.9	-1.7
Visual Genome	17.5	24.1	+6.6
TextVQA	5.4	8.8	+3.4
Total	100.0	100.0	0.0

Repair makes net additions from GQA, Visual Genome, and TextVQA, and net removals from COCO and OCR-VQA. COCO has the largest absolute exchange-set share change, at −15.5 percentage points. Since the exchanges affect 1,877 of $B = 93{,}884$ visual records, this corresponds to a subset-level change of approximately $-15.5 \times 1{,}877/93{,}884 = -0.31$ percentage points, consistent with the maximum source shift in Table 8. Thus the regional repair produces a modest change in the overall source mixture while keeping the subset size fixed.

Inserted records have a higher mean alignment percentile than removed records (0.756 versus 0.742), but a lower mean local-diversity percentile (0.624 versus 0.681). Their mean fused score before the final percentile transformation is therefore lower (0.6900 versus 0.7115), consistent with replacing records from the global top-B subset with records outside it.

D Sensitivity and Robustness

Repair has two tunable constraints: the edit fraction ϵ and the per-swap priority-loss bound τ. We study them with one-factor sensitivity analyses rather than a Cartesian grid:

$$\epsilon \in \{0.005, 0.01, 0.02, 0.05\} \quad \text{with } \tau = 0.05,$$

and

$$\tau \in \{0.01, 0.02, 0.05, 0.10\} \quad \text{with } \epsilon = 0.02.$$

The main comparisons use $(\epsilon, \tau) = (0.02, 0.05)$. The sweeps report selection-side behavior and downstream results for the trained settings; dashes indicate settings without a reported downstream score.

Table 13: **Sensitivity to edit fraction ϵ.** We fix $\tau = 0.05$.

ϵ	Swaps	Residual D	Reduction	Loss ($\times 10^{-4}$)	Retained	Mean
0.005	469	5,343	8.1%	1.08	99.50%	–
0.010	938	4,874	16.1%	2.08	99.00%	47.93
0.020	1,877	3,935	32.3%	4.39	98.00%	48.15
0.050	1,926	3,886	33.1%	4.56	97.95%	48.22

Increasing ϵ permits more replacements and therefore produces larger structural corrections, at the cost of greater deviation from the original top-B subset. At the default $\epsilon = 0.02$, Repair replaces 1,877 records, reduces the deficit by 32.3%, and retains 98.00% of the original selection.

Table 14: **Sensitivity to per-swap loss bound τ.** We fix $\epsilon = 0.02$.

τ	Swaps	Residual D	Reduction	Loss ($\times 10^{-4}$)	Mean
0.01	872	4,940	15.0%	0.82	–
0.02	1,672	4,140	28.8%	2.91	48.19
0.05	1,877	3,935	32.3%	4.39	48.15
0.10	1,877	3,935	32.3%	4.39	48.17

Tightening τ restricts the set of admissible replacements: at $\tau = 0.01$, only 872 swaps are accepted, compared with 1,877 at the default $\tau = 0.05$. Increasing the bound beyond 0.05 produces no further correction under the fixed edit budget, indicating that ϵ, rather than the per-swap loss constraint, becomes active in this regime.

For any setting, the construction guarantees

$$L_q(S_T) \leq \frac{T\tau}{B}.$$

With $B = 93{,}884$, the corresponding aggregate upper bounds for $\tau = \{0.01, 0.02, 0.05, 0.10\}$ are $\{0.929, 3.562, 9.996, 19.993\} \times 10^{-4}$, respectively. Table 14 reports the empirical aggregate loss and the trained downstream checks. The small downstream differences are single-run observations; the table does not measure training-seed variability at each setting.

D.1 Coverage-Partition Robustness

For each alternative, overlap is $100|S \cap S_{\text{default}}|/B$ between the repaired visual subsets. Each row varies one partition setting from the default of 128 clusters, three radial strata, and clustering seed 42.

Table 15: **Robustness to coverage-region construction.**

Axis	Setting	Initial D	Swaps	Reduction	Overlap	Mean
Visual clusters	64	5,326	1,877	35.2%	96.7%	47.96
	128	5,812	1,877	32.3%	100.0%	48.15
	256	6,284	1,877	29.9%	96.1%	–
Radial strata	1	5,401	1,663	30.8%	96.9%	–
	3	5,812	1,877	32.3%	100.0%	48.15
	5	5,994	1,877	31.3%	96.4%	–
Clustering seed	42	5,812	1,877	32.3%	100.0%	48.15
	123	5,865	1,877	32.0%	96.5%	–
	3407	5,721	1,877	32.8%	96.6%	–

Across these settings, Repair continues to reduce the deficit under alternative region constructions, while the repaired subsets are similar but not identical to the default selection. We retrain the 64-cluster alternative as a representative downstream check; the remaining alternatives are reported as selection-side robustness diagnostics.

D.2 Local-Diversity Robustness

Table 16 reports score correlations, selected-subset overlap, and downstream means as the neighborhood size varies. The default $k = 20$ provides the reference for correlation and overlap.

Table 16: **Robustness to neighborhood size.** Overlap is against default $k = 20$.

k	Spearman ρ	Top-B overlap	Mean
5	0.973	95.8%	46.62
10	0.988	97.4%	46.71
20	1.000	100.0%	47.18
50	0.982	96.9%	46.69

E Efficiency Details

Table 17 reports wall-clock time on a machine with an NVIDIA A800-SXM4-80GB GPU. RankRepair's feature preparation extracts only Layer 1 visual and instruction representations over the full candidate pool (approximately 3.0 hours). Cached selection includes loading these features, PCA/CCA

fitting, KNN computation, partition construction, scoring, index materialization, and writing the subset. These pipeline timings differ from the score-cached core operations below.

Table 17: **End-to-end selection efficiency at 100K.** Feature preparation includes the frozen-representation extraction required by each method. Cached selection includes selection and final subset construction from the cached representations.

Method	Feature preparation	Cached selection	End-to-end
PRISM	2.6 h	3.8 min	2.7 h
DataTailor	8.2 h	6.4 min	8.3 h
ScalSelect	3.4 h	4.7 min	3.5 h
Rank-only	3.0 h	2.1 min	3.0 h
RankRepair	3.0 h	3.5 min	3.1 h

Given cached final scores, global Rank-only top-B selection takes about 0.006 seconds. With the partition also cached, Repair adds about 1.34 seconds. These core-operation timings exclude feature extraction, PCA/CCA fitting, KNN computation, partition construction, and full subset materialization. Scores and the partition are reused across budgets and Repair-parameter sweeps. Repair changes the training subset without adding training or inference operations.

F Additional Model and Data-Pool Evaluation

Table 3 reports Qwen3-VL-4B trained on 57,000 records selected from a common Vision-Flan candidate pool of 186,103 records (30.63% retention), using the same budget for all methods. The pool is provided by `annotation_191-task_1k.json`. The RankRepair checkpoint is `mmselect-qwen_vf57k_mvr_top_s42`, with checkpoint identity `967bdb6915c5346379de7ea8a914056ad03217b3d563974cc09d838f5f6c14f9`. The table reports nine tasks with OCRBench divided by 10; MME is omitted, so the main study's ten-task Mean is not reported. This comparison evaluates RankRepair with a different model family and training pool from the main LLaVA study.