MATCHED COUNTERFACTUAL ATTRIBUTION: A CONTROLLED EVALUATION OF DATA SELECTION FOR VIDEO INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting instruction data is an increasingly important way to control multimodal post-training cost, yet reported gains are difficult to interpret when subset size, baseline choice, training variation, and evaluator behavior change together. We revisit this controlled policy-comparison setting with selector-specific matched counterfactuals, a frozen 10,656-example budget, and a crossed subset-draw by training-seed design. The design separates the total policy contrast ($S - U$), the shift into the realized data and processing envelope ($M - U$), and residual ranking within that envelope ($S - M$). We apply it to a frozen mixed image–video pool with Qwen3-VL-8B training and response-invariant evaluation on six benchmarks. In a homogeneous-H20 3-draw by 3-seed study, ScalSelect yields $+1.137$ percentage points (pp) for $S - U$ (95% CI $[+0.312, +1.963]$), but its $S - M$ estimate changes from -0.233 pp under three arbitrary diagonal pairs to $+0.263$ pp under the crossed design (CI $[-0.131, +0.657]$). Thus the pairing reverses the attribution estimate without establishing a positive ranking effect. A mechanism-distinct, homogeneous-A800 XMAS adaptation gives $+0.530$ pp for $S - U$ and -0.009 pp for $S - M$, with wide intervals spanning both signs. ScalSelect intervals use three training-seed marginals after averaging three control draws within seed; XMAS intervals use 5 marginals after averaging 5 draws within seed. The results do not rank or reject selectors in general. They show that a selector gain is interpretable only relative to an explicit counterfactual resolution, stochastic design, and training–evaluation contract.

1 INTRODUCTION

Data selection seeks to make multimodal instruction tuning more efficient by prioritizing examples for quality, diversity, visual relevance, or target utility (Yu et al., 2025; Ma et al., 2025; Yan et al., 2025; Dong et al., 2026b; Hong et al., 2026). Its success is often summarized by one contrast: performance on the selected subset minus performance on an equal-size random subset. This is a useful deployment comparison, but not necessarily an explanation of why the policy works. A selector can simultaneously change the proportions of sources and modalities, the processing lengths of retained examples, and their ordering within the resulting envelope. Moreover, drawing a new random subset and retraining a fixed subset are different interventions. Collapsing all of these choices into one contrast leaves its scientific attribution ambiguous.

We study this ambiguity as a counterfactual-attribution problem. Given a selected subset S, an unrestricted equal-size random subset U, and a selector-specific random subset M matched to the allocation and processing envelope realized by S, we distinguish the total policy contrast $S - U$, the realized-envelope contrast $M - U$, and residual within-envelope ranking $S - M$. The distinction is operational: every subset is materialized, trained, and evaluated under a frozen contract. We also cross random-subset draws with training seeds so that uncertainty in the counterfactual is not hidden inside uncertainty in optimization.

The decisive observation comes from a homogeneous-H20 3-draw by 3-training-seed study of ScalSelect (Wu et al., 2026). Its original diagonal pairing gives $S - M = -0.233$ percentage points (pp), whereas averaging the three counterfactual draws within each training seed gives $S - M =$

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.

$+0.263$ pp. The $+0.496$ pp shift reverses the sign of the estimate, although its seed-marginal interval still crosses zero. A complete homogeneous-A800 study of a one-frame Qwen3-VL adaptation of XMAS (Naharas et al., 2025) produces a different ladder: its matched point estimate is negative and its uncertainty is much wider. This is not a cross-selector or cross-hardware comparison. Instead, XMAS shows how the same reporting protocol describes a separately contracted selector study without requiring the numerical pattern to agree.

Our contribution is a reusable evaluation design, supported by two mechanism-distinct worked examples:

- We derive progressive random controls from each selector's realized output and report explicitly named total-policy, envelope, and residual ranking contrasts.
- We cross counterfactual subset draws with training seeds and show an attribution sign reversal not visible under the original diagonal pairing.
- We validate every admitted cell and expose the independent unit, hardware contract, and matching boundary, yielding a reporting template for bounded multimodal data-efficiency claims.

This is neither a new selector nor a verdict on data selection. It is an evaluation discipline that allows a policy to retain practical value while making clear what evidence, if any, supports ranking inside its realized data and resource envelope.

2 Related Work

Multimodal instruction-data selection. Recent selectors combine signals for visual necessity, informativeness, diversity, consistency, target influence, or representation coverage. TIVE allocates data across tasks before selecting examples within each task (Liu et al., 2024); DataTailor and MLLM-Selector combine quality- or necessity-oriented scores with diversity (Yu et al., 2025; Ma et al., 2025); and CoIDO and DICS couple sample utility or multimodal consistency with global diversity (Yan et al., 2025; Hong et al., 2026). ScalSelect instead uses instruction-attended visual representations and a dominant-subspace objective (Wu et al., 2026), while XMAS selects from cross-modal alignment trajectories observed during proxy training (Naharas et al., 2025). We use the latter two as mechanism-distinct policies, not as components of a new scoring rule.

Allocation, mixture, and targeting. Allocation is already an explicit optimization target. MaD-Mix and MixAtlas optimize modality, domain, or supervision mixtures (Xie et al., 2026; Wen et al., 2026), and DataComp-VLM contrasts mixture-preserving local filtering with global filtering that changes source proportions (Farina et al., 2026). Targeted methods including LESS, TAROT, SEED, and DataProphet use gradient, transport, graph, or transfer signals to operationalize downstream relevance (Xia et al., 2024; Feng et al., 2024; Zhang et al., 2026; Qi et al., 2026). Thus neither mixture awareness nor the distinction between allocation and ranking is new by itself. Our question is how to attribute a realized policy contrast after matching the particular envelope produced by that policy.

Controlled curation evaluation. Chasing Random evaluates instruction-selection strategies across sources, budgets, and tasks with repeated random trials and finds that conclusions can depend on the setup (Diddee & Ippolito, 2024). Fixed-protocol multimodal curation likewise compares random, difficulty, diversity, and mixture variants while exposing run-to-run variation (Shin et al., 2026). DataComp-VLM fixes model and training scales while comparing curation choices, and Curation-Bench fixes the model, recipe, and evaluation harness while reporting same-budget random comparisons (Farina et al., 2026; Kang et al., 2026). Repeated random baselines and fixed protocols are therefore important precedents, not our novelty. Our narrower addition is their joint use in a selector-specific progressive control ladder, with subset identity crossed against training randomness and with each resulting contrast assigned an explicit attribution target.

Scope relative to transfer claims. Selectors have also been evaluated across model families, scales, and data pools; Once-For-All explicitly trains a reusable selector for such transfer (Dong et al., 2026a). Our secondary model-stack condition has a different role: it probes whether a matched

contrast can be transported beyond the stack on which it was measured. It neither isolates a backbone effect nor introduces a transferable selector.

3 MATCHED-COUNTERFACTUAL ATTRIBUTION

The evaluated object is a realized policy. A selector determines more than a score ordering. Its observation rule and selection procedure may change which candidates are eligible, how the budget is allocated across sources or capability groups, and the processing-cost profile of retained examples. Comparing the selected subset with uniform sampling from the full pool assigns all of these changes to the named policy. That total contrast can answer a deployment question, but it does not isolate ranking among otherwise comparable candidates.

Let S denote the selected subset, U_d the dth unrestricted equal-size random subset, and M_d the dth random subset matched to an envelope realized by S. For downstream score $Y(A;k)$ from training subset A with seed k, define the control-draw marginal

$$\overline{Y}_A(k) = \frac{1}{D}\sum_{d=1}^{D} Y(A_d;k), \qquad A \in \{U, C, M\}, \tag{1}$$

where C is used when the selector-specific ladder includes an intermediate composition-matched control. We estimate three contrasts by averaging paired seed marginals:

$$\widehat{\tau}_{S-U} = \frac{1}{K}\sum_{k}\left[Y(S;k) - \overline{Y}_U(k)\right], \qquad \text{total policy contrast,} \tag{2}$$

$$\widehat{\tau}_{M-U} = \frac{1}{K}\sum_{k}\left[\overline{Y}_M(k) - \overline{Y}_U(k)\right], \qquad \text{realized-envelope contrast,} \tag{3}$$

$$\widehat{\tau}_{S-M} = \frac{1}{K}\sum_{k}\left[Y(S;k) - \overline{Y}_M(k)\right], \qquad \text{within-envelope ranking contrast.} \tag{4}$$

The identity $\widehat{\tau}_{S-U} = \widehat{\tau}_{M-U} + \widehat{\tau}_{S-M}$ is exact for the declared finite design. It is an accounting identity and does not identify a mediating mechanism: unmeasured differences between subsets may remain, and none of the three terms is an intrinsic selector property.

Counterfactual resolution. We use *counterfactual resolution* for the factors that a random control must preserve. The coarsest control U fixes only the example budget. A composition control C additionally reproduces the selected source and temporal quotas. The most resolved control M also matches joint processor-cost bins and constrains aggregate token totals where feasible. For this four-level ladder,

$$\widehat{\tau}_{S-U} = \widehat{\tau}_{C-U} + \widehat{\tau}_{M-C} + \widehat{\tau}_{S-M}. \tag{5}$$

Each additional constraint conditions the residual contrast on another set of matched, observed factors; it neither identifies a causal component nor guarantees that the point estimate or its uncertainty shrinks. Because different selectors realize different envelopes, the matching rules must be selector-specific and disclosed explicitly.

Table 1 makes the claim–contrast correspondence operational. The rows are not competing metrics: they answer different questions, and a narrower contrast cannot retroactively change the meaning of a broader one.

Subset draws and training seeds are distinct axes. The $D \times K$ control cells are not DK independent repetitions: selected outcome $Y(S;k)$ is shared by every control draw at seed k, and the same subset draw is reused across seeds. Pairing only $(d_1, k_1), (d_2, k_2), \ldots$ therefore makes the attribution depend on an arbitrary diagonal through the grid. Our estimator instead averages all declared draws within each seed and uses the K seed marginals as the independent units for its Student-t interval. Cell-level values remain visible as finite-design diagnostics.

More explicitly, when $D = K$, a diagonal selected by a bijection π estimates

$$\widehat{\tau}^{\,\mathrm{diag}}_{S-M}(\pi) = \frac{1}{K}\sum_{k}\left[Y(S;k) - Y(M_{\pi(k)};k)\right], \tag{6}$$

Table 1: Claims licensed by each counterfactual contrast. "Required" means required for that attribution target, not sufficient for causal identification.

Claim target	Contrast	Information that remains part of the contract
Deploy the complete realized policy	$S - U$	Candidate universe, budget, observation, allocation, ranking, and processing differences
Locate the observed envelope shift	$M - U$	Exact matching constraints and all attributes that remain unmatched
Assess ranking inside that envelope	$S - M$	Residual content differences, subset draws, training seeds, and evaluation stack
Claim transport or efficiency	Repeat the ladder	New pool/model contracts; measured selection and training cost for efficiency claims

whereas Eq. 4 averages every declared matched draw at every seed. The diagonal quantity can vary with π even though the selected subset, seeds, and matching rule are fixed. In a balanced grid, averaging Eq. 6 over all $K!$ bijections recovers the crossed estimate exactly. Crossing therefore removes dependence on an arbitrary draw–seed alignment within the observed design; it does not create DK independent replications or estimate variation beyond the declared draws and seeds.

The resulting question is deliberately narrower than whether a selector "works": *which part of a realized policy is supported by a specified comparison, under which data, model, training, and evaluation contract?*

4 Experimental Protocol

4.1 Candidate pool, selectors, and controls

We build a frozen heterogeneous image–video instruction pool with 106,596 stable candidate identities. Model-specific preprocessing is run before selection; over-length or failed examples are excluded explicitly rather than silently dropped during training. Qwen3-VL-8B admits 84,783 candidates in the primary ScalSelect condition and 84,682 in the one-frame XMAS condition. The common Qwen/InternVL transfer universe contains 60,317 raw identities. Every reported subset contains unique identities and is frozen by SHA-256 before training.

We evaluate two recent, mechanism-distinct selectors. ScalSelect is a training-free method based on instruction-attended visual representations and a dominant-subspace objective (Wu et al., 2026). We use its official importance-scoring path and deterministic ID tie breaking. XMAS measures the training trajectory of cross-modal attention alignment in a proxy model, clusters those trajectories, and samples stable examples with balanced cluster allocation (Naharas et al., 2025). Our condition implements the core of XMAS Algorithm 1 with a Qwen3-VL architecture adapter. Images are unchanged, while each video is represented to the *selector* by one deterministic midpoint frame; downstream training still receives the original example. This is a one-frame selector-observation condition, not an unmodified or video-native XMAS reproduction. We additionally retain the reverse Q/K orientation as an implementation-sensitivity control because the released function signature and positional call differ.

Unrestricted controls sample from the full admitted universe at fixed example budget. Composition-matched controls reproduce exact quotas over source domain, source dataset, and temporal category. Processor-cost controls refine these cells by the joint empirical-quartile bins of full-sequence, visual, and loss-bearing token counts. The ScalSelect matched controls additionally keep each aggregate token total within 1% of the selected subset. This is exact joint-bin matching, not per-example exact-token or measured-FLOP matching.

4.2 Training and evaluation

The primary target is Qwen3-VL-8B-Instruct (Bai et al., 2025); a secondary model-stack condition uses InternVL3-8B (Zhu et al., 2025). We full-fine-tune for one epoch with global batch size 32, peak learning rate 2×10^{-5}, cosine decay, 0.1 warm-up ratio, bfloat16, and a maximum length of 8,192

Table 2: Primary crossed contracts. Here S, U, C, and M denote the selected, unrestricted, composition-matched, and processor/token-matched arms. ScalSelect intervals use three training-seed marginals after averaging three random-control draws within seed; XMAS intervals use 5 marginals after averaging 5 draws. Cells within each grid are dependent and are not counted as independent repetitions.

Selector	Model	Budget	Training cells	Independent unit	Hardware
ScalSelect	Qwen3-VL-8B	10,656	$3S + 9U + 9M = 21$	$n = 3$ seed marginals	H20
XMAS adapter	Qwen3-VL-8B	10,656	$5S + 25U + 25C + 25M = 80$	$n = 5$ seed marginals	A800

tokens. Each job uses 32 accelerators. The primary budget is 10,656 examples (333 optimizer steps); an appendix-only design-sensitivity screen uses 26,656 examples (833 steps). Selected and control arms at the same seed share all optimization settings.

We evaluate MVBench, Video-MME, MLVU, LVBench, MMMU, and RealWorldQA (Li et al., 2024; Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025; Yue et al., 2024; xAI, 2024). The primary aggregate is the unweighted mean of the six benchmark accuracies; it does not weight by benchmark size. We also report video and image macros and all benchmark-level effects in the appendix. A response-invariant multiple-choice normalizer is applied uniformly to every arm, with raw responses retained.

4.3 Uncertainty and admission

For ScalSelect, we first average three control draws within each training seed; for XMAS, we average 5 draws. We then form a paired contrast with the selected outcome at that seed. Figures and main-text intervals report two-sided Student-t intervals over three ScalSelect seed marginals and 5 XMAS seed marginals. These intervals are low-powered, and crossing zero is not evidence of equivalence. A product bootstrap that resamples the two observed axes is retained only as a finite-grid sensitivity analysis in the appendix; it does not increase the independent sample size.

The crossed completions were designed after observing earlier results and are therefore retrospective, non-preregistered analyses. ScalSelect retains its original three training seeds while completing the control-draw–training-seed grid. XMAS atomically replaces a complete three-seed fallback with a complete five-seed, five-draw precision extension under an outcome-independent admission rule. Our claims concern these finite designs and not a population error rate for diagonal pairing.

A run enters the evidence set only if its platform record is terminal, every distributed rank ledger reaches the required stages, the unique final checkpoint exists, all six benchmark result and sample counts match the contract, and recorded hashes validate. Failed and migrated jobs remain in the provenance ledger. If training completed but evaluation failed, only the frozen checkpoint is re-evaluated; it is never counted as a new training replicate. Hardware is homogeneous within each primary selector study and is never pooled across the H20 and A800 blocks.

5 Results

5.1 A diagonal pairing reverses the attribution estimate

The original ScalSelect comparison paired one matched-control draw with each training seed. Its three diagonal $S - M$ contrasts average -0.233 percentage points (pp). Completing the Cartesian design changes the comparison: after averaging all three matched draws within each seed, the seed marginals are $+0.404$, $+0.293$, and $+0.092$ pp. Their mean is $+0.263$ pp, with a two-sided 95% Student-t interval of $[-0.131, +0.657]$. The crossed estimate is therefore $+0.496$ pp higher and has the opposite sign (Fig. 1). Across all six possible one-to-one pairings of three draws with three seeds, the descriptive means range from -0.233 to $+0.878$ pp and five are nonnegative; their average is exactly the crossed estimate, $+0.263$ pp. The crossed estimate uses every draw at every seed and is therefore invariant to this arbitrary pairing choice.

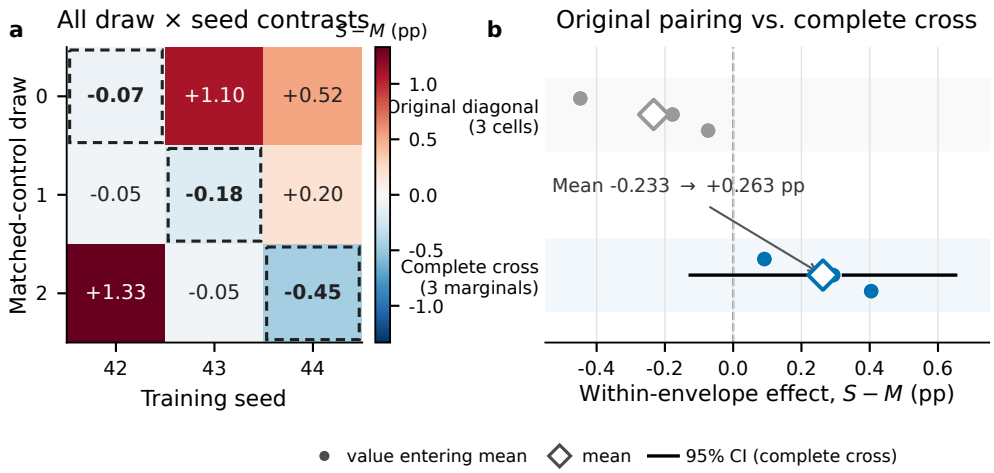

Figure 1: **Pairing the two stochastic axes changes attribution.** **a**, All nine ScalSelect $S - M$ contrasts from three matched-subset draws and three training seeds; the dashed diagonal marks the three pairs in the original estimate. Cells are dependent because each selected outcome is shared across the three draws at a seed. **b**, Diagonal-only pairs give -0.233 pp, whereas the complete-cross estimator gives $+0.263$ pp. Filled circles are values entering each displayed mean, open diamonds are means, the solid line is the crossed estimator's 95% Student-t interval ($[-0.131, +0.657]$), and the vertical dashed line marks zero effect. The result concerns the sign of the finite-design estimate, not the sign of a population ranking effect.

The selected runs and training seeds are unchanged. Completion adds six off-diagonal control cells, evaluating every draw at every seed. Four of nine cell contrasts are positive and five are negative, making any single diagonal contingent on the sampled controls. The three-seed interval still spans zero. The result therefore establishes finite-design sensitivity, not a positive ranking benefit; completion changes comparisons, not replication.

5.2 Counterfactual resolution changes what receives credit

The complete ScalSelect cross yields a total policy contrast of $+1.137$ pp against unrestricted random (95% CI $[+0.312, +1.963]$). Moving from unrestricted controls to the selected policy's composition and processor-cost envelope gives $M - U = +0.875$ pp (CI $[-0.134, +1.883]$), while ranking within that envelope gives $S - M = +0.263$ pp (CI $[-0.131, +0.657]$). The means satisfy $S - U = (M - U) + (S - M)$ exactly. This says where the finite-policy contrast is located under the declared control ladder; it does not make $M - U$ a causal mediator. The matched control fixes source and temporal quotas, joint processor-cost bins, and aggregate-token tolerances, but not latent content, measured FLOPs, wall-clock time, or examples one by one.

The mechanism-distinct XMAS adapter gives a different pattern (Fig. 2). Its total policy contrast is $+0.530$ pp (CI $[-0.748, +1.807]$), the combined realized-envelope contrast is $M - U = +0.539$ pp (CI $[-0.424, +1.502]$), and the residual ranking contrast is $S - M = -0.009$ pp (CI $[-0.810, +0.792]$). At the finer resolution, $C - U = +0.275$ pp (CI $[-0.564, +1.113]$) and $M - C = +0.264$ pp (CI $[-0.026, +0.554]$). Both accounting identities close to numerical precision. The XMAS intervals are wide and all span zero, so they support neither superiority, inferiority, nor equivalence. They also show that matching changes the estimand but need not monotonically reduce its magnitude.

5.3 Task-level and candidate-pool checks bound generalization

The aggregate contrasts conceal substantial task heterogeneity (Table 3). ScalSelect's residual $S - M$ point estimate is positive on five benchmarks and negative on one; XMAS is positive on three and

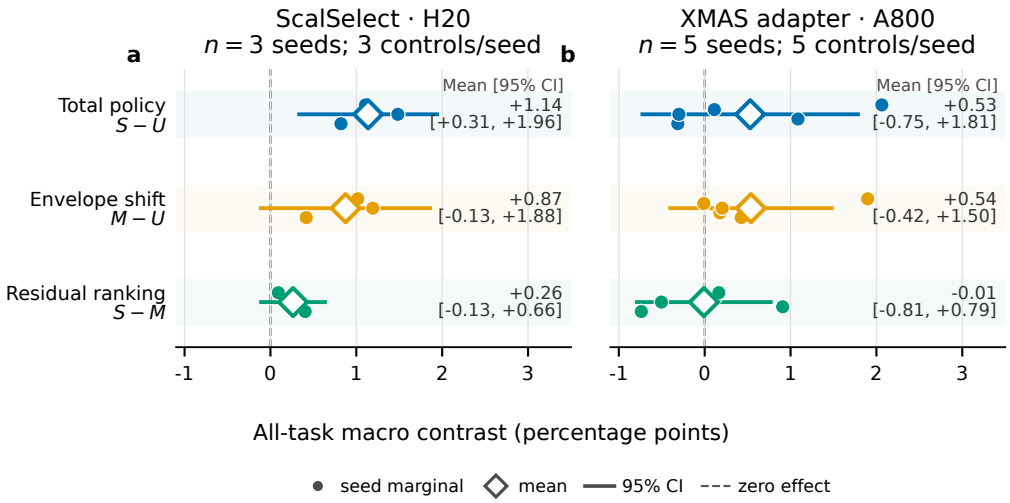

Figure 2: **The same labels denote contract-conditional contrasts.** Filled circles are training-seed marginals after averaging the declared counterfactual draws within that seed. Diamonds are means and horizontal lines are two-sided 95% Student-t intervals; the dashed vertical line is zero effect. ScalSelect uses $n = 3$ seed marginals and three draws per seed; XMAS uses $n = 5$ seed marginals and 5 draws per seed. ScalSelect is evaluated on homogeneous H20 hardware; the XMAS adapter is evaluated separately on homogeneous A800 hardware, and the two blocks are not pooled. $S - U$ measures the complete realized policy, $M - U$ the shift to the matched envelope, and $S - M$ residual ranking within that envelope.

negative on three. Eleven of the twelve task-level residual intervals span zero. The remaining interval, for ScalSelect on MLVU, is one unadjusted secondary comparison among twelve and does not establish a task-general ranking advantage. We therefore use the prespecified six-task macro for the primary attribution statement and retain the disaggregated results to show where that macro does and does not transport.

A second deterministic ScalSelect candidate-pool sample from the same source reservoir provides a further scope check. Its complete 21-cell A800 grid gives $S - U = +0.835$ pp (95% CI $[-0.278, +1.948]$), $M - U = +0.513$ pp (CI $[-0.613, +1.639]$), and $S - M = +0.322$ pp (CI $[+0.103, +0.541]$). The positive residual interval does not overturn the attribution result: Pool-2 changes both the sampled candidate pool and the hardware contract, so its difference from Pool-1

Table 3: Benchmark-level primary contrasts, reported as mean [two-sided 95% Student-t interval] in percentage points. ScalSelect estimates use $n = 3$ training-seed marginals after averaging three draws within seed; XMAS estimates use $n = 5$ marginals after averaging 5 draws. These unadjusted, low-powered benchmark intervals are descriptive; the all-task macro is the primary outcome.

Benchmark	ScalSelect / H20 ($n = 3$)		XMAS adapter / A800 ($n = 5$)	
	Total $S - U$	Ranking $S - M$	Total $S - U$	Ranking $S - M$
MVBench	$+2.13\,[-0.46, +4.73]$	$+0.49\,[-4.04, +5.02]$	$+0.83\,[-1.43, +3.10]$	$-0.10\,[-1.98, +1.78]$
Video-MME	$+0.63\,[-0.12, +1.38]$	$+0.32\,[-0.77, +1.41]$	$+0.14\,[-1.35, +1.64]$	$+0.26\,[-0.80, +1.33]$
MLVU	$+1.87\,[-0.93, +4.67]$	$+1.04\,[+0.40, +1.69]$	$+2.08\,[-0.84, +5.00]$	$+1.05\,[-0.78, +2.87]$
LVBench	$-0.13\,[-2.03, +1.77]$	$-0.65\,[-3.33, +2.02]$	$-0.30\,[-2.08, +1.47]$	$-0.63\,[-1.75, +0.49]$
MMMU	$+1.26\,[-0.88, +3.40]$	$+0.17\,[-3.88, +4.23]$	$-0.33\,[-1.84, +1.19]$	$-0.70\,[-1.88, +0.48]$
RealWorldQA	$+1.06\,[-1.05, +3.17]$	$+0.20\,[-0.46, +0.86]$	$+0.75\,[-0.92, +2.43]$	$+0.07\,[-1.50, +1.64]$

cannot be assigned to either factor and the estimates are not pooled. It instead shows why candidate-pool identity belongs in the contract attached to a selector claim. Full construction and admission details appear in Appendix F.

5.4 THE STABLE CONCLUSION IS ABOUT ATTRIBUTION, NOT SELECTOR QUALITY

The two selector studies do not exhibit a universal effect pattern, nor were they designed to rank ScalSelect against XMAS. What survives their different mechanisms and hardware-homogeneous contracts is more limited: an equal-size unrestricted control assigns the entire realized policy to the selector; matching changes the scientific target of the contrast; and retraining one control subset cannot quantify uncertainty over which admissible control was drawn. Task-level heterogeneity and the Pool-2 scope check reinforce that the numerical estimate does not detach from its contract. Secondary model-stack, budget, and hardware diagnostics in the appendix further bound transport. We therefore interpret every reported value as conditional on its selector, candidate pool, counterfactual resolution, model stack, and evaluation suite.

6 DISCUSSION

What the results change. Method papers often use "selector gain" for both a practical deployment contrast and an explanation of where that contrast comes from. Our results separate these uses. $S - U$ is appropriate when a practitioner adopts the entire policy, including its allocation and processing consequences. $S - M$ is appropriate when the scientific claim concerns ranking inside the realized envelope. Reporting both avoids a false choice between celebrating the total gain and dismissing the selector: a useful policy contrast may be positive even when the available evidence does not isolate positive within-envelope ranking.

This distinction also changes how negative or uncertain estimates should be read. The crossed ScalSelect ranking interval and every XMAS interval contain values on both sides of zero; none establishes equivalence. Conversely, the positive ScalSelect total-policy interval does not by itself identify ranking as its source. Matching sharpens the question being asked; it is neither a procedure for explaining away gains nor a guarantee that an estimate shrinks.

Why the counterfactual is selector-specific. There is no universal matched random baseline. A clustering policy can impose cluster allocation; a subspace policy can alter modality, source, and token profiles; a learned policy can change what it observes during scoring. The counterfactual must therefore be derived from the realized output and disclosed as a list of constraints. An equal-size unrestricted random result remains a valid total-policy comparison, but it should not automatically receive a mechanistic interpretation.

Implications for experimental design. Subset redraws and training seeds are crossed factors, not interchangeable replicates. Repeatedly training one random subset can characterize optimization variability conditional on that subset, but cannot characterize the distribution of admissible counterfactuals. Conversely, training many draws only once leaves their outcomes entangled with optimization. A compact attribution study should therefore prefer a balanced rectangle and report seed marginals after integrating over declared draws. Every interval should name its independent unit; the cells used to construct a marginal are not extra independent replications.

The grid should be planned around the estimand rather than expanded by adding nominally different runs. With K training seeds and D draws for each control family, a selected arm plus unrestricted and matched controls requires $K + 2DK$ training cells. Adding a composition control requires another DK cells. This accounting makes the tradeoff explicit: increasing K strengthens the number of seed marginals available for interval estimation, whereas increasing D resolves sensitivity to the sampled counterfactual within each seed. Neither substitutes for the other, and neither licenses inference over unobserved candidate pools. We do not infer that three or five levels are universally sufficient from the finite designs studied here.

The 26,656-example stress test illustrates this non-substitutability. One-draw/one-seed subdesigns disagree with the complete-grid sign in 27.2% of cases and with its 0.5-pp decision in 45.1%. Retaining all seeds but one draw reduces these rates to 13.0% and 33.3%; retaining all draws but one

seed gives 25.9% and 46.3%. These dependent, mixed-hardware enumerations are not universal error rates. They show only that neither stochastic axis substitutes for the other.

A minimum attribution record. A reproducible selector claim should jointly disclose the candidate universe, realized policy, preserved and unresolved control factors, crossed draw–seed matrix, independent unit, and training–evaluation contract. Total and matched contrasts should be reported with uncertainty and disaggregated outcomes. This turns "better than random" into a checkable statement and identifies a new pool, model, or evaluator as a new contract unless the relevant axes are fixed.

Retrospective diagnosis and prospective use. Because our crossed completions followed earlier results, the sign change diagnoses this finite design rather than a preregistered error rate. A prospective study should freeze the universe, aggregate, controls, stochastic levels, admission gates, and estimator, then admit the complete rectangle regardless of direction. Retrospective use can expose dependence on one draw–seed alignment but cannot recover a design's prospective operating characteristics.

A hierarchy of transport evidence. The evidence has three levels. The crossed estimator characterizes one frozen contract; benchmark contrasts expose heterogeneity within its trained models; and Pool-2, model-stack, budget, and hardware diagnostics alter contract axes. The latter are boundary tests: agreement does not prove invariance, and disagreement cannot identify the responsible axis. Large experiment suites should therefore designate one estimation contract, use disaggregated outcomes to expose hidden heterogeneity, and treat changed-contract runs as hypotheses rather than pooling everything into one robustness tally.

Limitations. ScalSelect has only three independent training-seed marginals, and XMAS has only 5; both studies are consequently low-powered. The XMAS adapter gives each video one midpoint frame during selection, so it is not an unmodified video-native reproduction. The ScalSelect control matches joint processor-cost bins and aggregate token totals, not examples, latent content, measured FLOPs, wall-clock cost, or selector overhead. The two primary blocks also use different accelerator types and are analyzed separately. Secondary InternVL, budget, and hardware diagnostics change multiple contract variables or use finite observed grids, so they bound interpretation rather than proving transfer. Finally, two selectors, one underlying source reservoir, and six benchmarks do not establish population-level generality.

Future studies can preregister a hierarchy of counterfactual resolutions, increase both stochastic axes, and test transport across candidate pools and training budgets. A stochastic selector adds a third randomness axis. Cost claims should also include selection overhead and realized training FLOPs rather than treating equal example counts as equal compute.

7 Conclusion

A multimodal data-selection gain is a contrast under a specified contract, not a scalar property of a selector. In the ScalSelect study, replacing an arbitrary diagonal by the crossed subset-draw–training-seed estimator reverses the sign of within-envelope attribution in the completed finite design. The mechanism-distinct XMAS study applies the same reporting protocol under a separate hardware and selector contract; it is not evidence for a cross-selector ordering or a shared effect sign. The practical prescription is simple: report the total policy contrast, derive matched controls from the selector's realized envelope, and keep counterfactual draws distinct from training seeds. This protocol does not decide whether data selection works in general; it makes bounded data-efficiency claims more interpretable and reproducible.

Reproducibility Statement

The appendix specifies subset construction, selector adaptations, training, evaluation, estimands, and statistical units. The accompanying anonymized artifact contains frozen subset and execution manifests, source hashes, strict validators, per-pair summaries, and scripts that regenerate every data-dependent paper-facing figure and generated result table. Failed, stopped, recovered, and superseded jobs are retained in a separate provenance ledger and are never counted as scientific repetitions.

AI Use Statement

Generative AI tools were used primarily to check grammar and improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take responsibility for the final manuscript.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Harshita Diddee and Daphne Ippolito. Chasing random: Instruction selection strategies fail to generalize. *arXiv preprint arXiv:2410.15225*, 2024. URL https://arxiv.org/abs/2410.15225.

Mingkang Dong, Hongyi Cai, Xiwen Lei, Jie Li, Tao Zhang, and Muxin Pu. Once-for-all: A train-once and select-anytime framework for multimodal instruction tuning. *arXiv preprint arXiv:2605.26761*, 2026a. URL https://arxiv.org/abs/2605.26761.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. VisNec: Measuring and leveraging visual necessity for multimodal instruction tuning. In *European Conference on Computer Vision*, 2026b. doi: 10.1007/978-3-032-37577-3_29. URL https://doi.org/10.1007/978-3-032-37577-3_29.

Matteo Farina, Vishaal Udandarao, Thao Nguyen, Selim Kuzucu, Maximilian Böther, et al. DataComp-VLM: Improved open datasets for vision-language models. *arXiv preprint arXiv:2606.28551*, 2026. URL https://arxiv.org/abs/2606.28551.

Lan Feng, Fan Nie, Yuejiang Liu, and Alexandre Alahi. TAROT: Targeted data selection via optimal transport. *arXiv preprint arXiv:2412.00420*, 2024. URL https://arxiv.org/abs/2412.00420.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://doi.org/10.1109/CVPR52734.2025.02245.

Yuyang Hong, Jinhui Guo, Jiaqi Gu, Lubin Fan, Ruixiang Wang, Kun Ding, Yue Wu, Shiming Xiang, and Jieping Ye. DICS: Exploring data intrinsic consistency for visual instruction selection. *arXiv preprint arXiv:2608.30209*, 2026. URL https://arxiv.org/abs/2608.30209.

Feiyang Kang, Hanze Li, Adam Nguyen, Mahavir Dabas, Jiaqi W. Ma, Frederic Sala, Dawn Song, and Ruoxi Jia. Can generalist agents automate data curation? *arXiv preprint arXiv:2606.04261*, 2026. URL https://arxiv.org/abs/2606.04261.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

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. *arXiv preprint arXiv:2403.09559*, 2024. URL https://arxiv.org/abs/2403.09559.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Jiayi Ji, Jie Lou, Debing Zhang, and Rongrong Ji. MLLM-Selector: Necessity and diversity-driven high-value data selection for enhanced visual instruction tuning. *arXiv preprint arXiv:2503.20502*, 2025. URL https://arxiv.org/abs/2503.20502.

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

Xuan Qi, Luxi He, Dan Roth, and Xingyu Fu. DataProphet: Demystifying supervision data generalization in multimodal LLMs. *arXiv preprint arXiv:2603.19688*, 2026. URL https://arxiv.org/abs/2603.19688.

Yosub Shin, Michael Buriek, Boris Sobolev, Pavel Bushuyeu, Vikas Kumar, Haoyang Xu, Samuel Watson, and Igor Molybog. Data-efficient curation for multimodal reasoning under fixed training protocols. *arXiv preprint arXiv:2601.10922*, 2026. URL https://arxiv.org/abs/2601.10922.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.02131. URL https://doi.org/10.1109/ICCV51701.2025.02131.

Bingbing Wen, Sirajul Salekin, Feiyang Kang, Bill Howe, Lucy Lu Wang, Javier Movellan, and Manjot Bilkhu. MixAtlas: Uncertainty-aware data mixture optimization for multimodal LLM midtraining. *arXiv preprint arXiv:2604.14198*, 2026. URL https://arxiv.org/abs/2604.14198.

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

xAI. Grok-1.5 Vision Preview. https://x.ai/news/grok-1.5v, 2024. Introduces the RealWorldQA benchmark.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Wanyun Xie, Francesco Tonin, and Volkan Cevher. MaD-Mix: Multi-modal data mixtures via latent space coupling for vision-language model training. *arXiv preprint arXiv:2602.07790*, 2026. URL https://arxiv.org/abs/2602.07790.

Yichen Yan, Ming Zhong, Qi Zhu, Xiaoling Gu, Jinpeng Chen, and Huan Li. CoIDO: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-5569. URL https://doi.org/10.52202/085713-5569.

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*, 2025. doi: 10.1109/ICCV51701.2025.00022. URL https://doi.org/10.1109/ICCV51701.2025.00022.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 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*, 2024. URL https://arxiv.org/abs/2311.16502.

Yuan Zhang, Lifeng Guo, Junwen Pan, Wenzhao Zheng, Wen Zhou, Kuan Cheng, Kurt Keutzer, and Shanghang Zhang. SEED: Targeted data selection by weighted independent set. *arXiv preprint arXiv:2605.15691*, 2026. URL https://arxiv.org/abs/2605.15691.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01278. URL https://doi.org/10.1109/CVPR52734.2025.01278.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Detailed attribution results

Table 4 gives the macro estimates underlying Figure 2. ScalSelect rows use three training-seed marginals after averaging three control draws within seed. XMAS rows use 5 seed marginals after averaging 5 draws within seed. The draw-by-seed cells are inputs to these marginals, not independent repetitions.

Table 4: All-task macro attribution contrasts. Intervals are two-sided 95% Student-t intervals over seed marginals after averaging control draws within seed. ScalSelect uses $n = 3$ seeds and three draws; XMAS uses $n = 5$ seeds and 5 draws. Values are percentage points.

Selector	Contrast	Mean	Seeds	95% CI
ScalSelect	$S - U$ (total policy)	$+1.137$	3	$[+0.312, +1.963]$
ScalSelect	$M - U$ (realized envelope)	$+0.875$	3	$[-0.134, +1.883]$
ScalSelect	$S - M$ (within-envelope ranking)	$+0.263$	3	$[-0.131, +0.657]$
XMAS adapter	$S - U$ (total policy)	$+0.530$	5	$[-0.748, +1.807]$
XMAS adapter	$C - U$ (composition increment)	$+0.275$	5	$[-0.564, +1.113]$
XMAS adapter	$M - C$ (processor/token increment)	$+0.264$	5	$[-0.026, +0.554]$
XMAS adapter	$M - U$ (realized envelope)	$+0.539$	5	$[-0.424, +1.502]$
XMAS adapter	$S - M$ (within-envelope ranking)	-0.009	5	$[-0.810, +0.792]$

The crossed ScalSelect completion and the XMAS 5-draw by 5-seed study are post-result attribution extensions, not preregistered experiments. Both are nevertheless complete under their frozen within-study homogeneous hardware contracts; all 21 ScalSelect and 80 XMAS physical cells pass the terminal-state, checkpoint, benchmark-count, and hash gates. No partial extension or historical cross-hardware pair contributes to Table 4.

B Subset and counterfactual construction

Candidate rows receive stable identifiers before any selector is applied. Random priorities are SHA-256 digests of a frozen namespace, draw seed, and stable identifier. Subset construction is deterministic conditional on the manifest. All arms contain exactly the declared budget and have no duplicate stable identifiers.

Composition matching fixes the Cartesian quotas of source domain, source dataset, and temporal category realized by the selected subset. The processor-cost refinement assigns each admitted candidate to joint empirical quartile bins of full-sequence, visual, and loss-bearing tokens, and fixes those joint-bin quotas inside every composition stratum. Aggregate token differences are recorded and, for the primary ScalSelect decomposition, are at most 1%. This procedure should not be described as per-example token matching, exact compute matching, or FLOP matching.

The unrestricted, matched, and selected subsets are constructed without using any downstream benchmark outcome. Control draws and training seeds are registered in the execution plan before the corresponding result is admitted. The original examples, rather than selector-only one-frame views, are used for downstream training.

C SELECTOR IMPLEMENTATION BOUNDARIES

ScalSelect. The official scoring path computes instruction-conditioned visual representations and applies the released dominant-subspace importance rule. Scores cover the complete declared candidate pool. We add only deterministic identifier tie breaking and the counterfactual builders; neither changes a strict score ordering.

XMAS. The adapter preserves Algorithm 1's seven proxy checkpoints, layer-summed text-query/visual-key attention, sum of the five leading singular values, trajectory variation, 1,000-cluster FAISS K-means, and balanced stable sampling. Qwen3-VL requires processor-derived text and visual positions in place of the released TinyLLaVA offsets. The paper-defined query/key orientation is used, while the alternative positional-call orientation is retained as a sensitivity arm. Each video is represented by a deterministic midpoint frame during selection. We therefore call the condition "XMAS Algorithm 1 with a Qwen3-VL architecture adapter," not an unmodified or video-native reproduction.

D TRAINING AND EVALUATION DETAILS

All target runs use one epoch, global batch size 32, per-device batch size 1, no gradient accumulation, full-parameter optimization with AdamW, peak learning rate 2×10^{-5}, cosine decay, warm-up ratio 0.1, weight decay 0.01, gradient clipping at 1.0, bfloat16, gradient checkpointing, and an 8,192-token maximum length. Distributed jobs use four nodes with eight GPUs per node. Seeds are applied to training and deterministic data ordering as declared by each plan.

The evaluation suite contains 4,000 MVBench items, 2,700 Video-MME items, 502 MLVU items, 1,549 LVBench items, 900 MMMU items, and 765 RealWorldQA items. Every admitted comparison uses identical item-key hashes across its arms. The all-task macro is the arithmetic mean of these six accuracies; the video macro averages the first four and the image macro averages the last two.

Legacy MLVU measurement audit. The registered policy study also exposed a response-format interaction that is separate from the crossed attribution results. A legacy parser accepted parenthesized option letters but rejected semantically equivalent responses such as `A.`. Applying one response-invariant normalizer to all retained predictions changed the largest observed score by 43.44 pp (499 of 502 rows) and changed the apparent policy ordering. This is a bounded diagnostic of these archived MLVU predictions, not evidence that all multiple-choice evaluators fail. The primary analyses use the same normalizer for every arm and retain raw responses for audit.

E STATISTICAL DETAILS

The primary intervals first average control draws within each training seed and then treat the paired seed marginals as independent units. A two-sided Student-t interval is shown because the selected and control arms share the training seed. ScalSelect contributes three seed marginals; XMAS contributes 5. Both designs remain low-powered. Their draw-by-seed contrasts are dependent and are never counted as independent observations.

As a sensitivity analysis, a product bootstrap independently resamples the observed subset-draw and training-seed levels, reconstructs their Cartesian product, and recomputes the mean. For ScalSelect, its 50,000-replicate intervals are $[+0.671, +1.801]$ pp for $S - U$, $[+0.210, +1.736]$ pp for $M - U$, and $[-0.179, +0.720]$ pp for $S - M$. These intervals describe resampling of the observed finite axes; they are not population intervals and do not increase $n = 3$.

The separate subdesign analysis enumerates all rectangular choices of one to three draw levels and one to three seed levels from an observed grid. Sign disagreement compares a subdesign mean with its complete-grid mean. Practical decision disagreement additionally compares whether the absolute mean exceeds a declared threshold of 0.5 pp in the same direction.

F FINITE-GRID AND SCOPE DIAGNOSTICS

Reduced-design sensitivity. At budget 26,656, a mixed-hardware finite grid yields 27.2% sign disagreement and 45.1% 0.5-pp decision disagreement for one-draw/one-seed subdesigns. The corresponding rates are 13.0%/33.3% for one draw and three seeds, and 25.9%/46.3% for three draws and one seed. These rates pool enumerated subdesigns across 18 outcome-by-policy groups. They are neither independent experiments nor estimates of a population error rate.

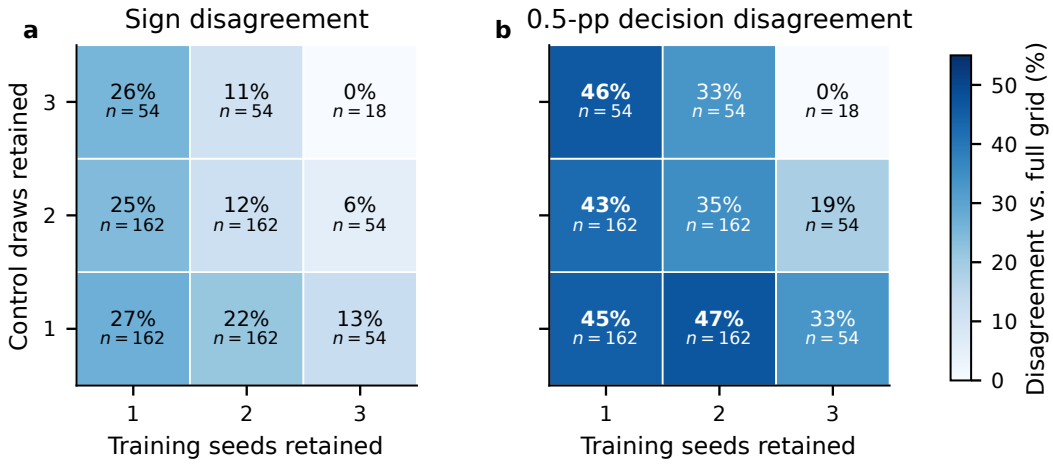

Figure 3: **Finite-grid sensitivity to reduced replication.** Each cell enumerates rectangular subdesigns of one observed three-draw by three-seed ScalSelect grid at budget 26,656. Percentages are disagreements with the corresponding complete-grid decision, pooled across 18 outcome-by-policy groups. Each cell reports the disagreement rate followed by n, the number of dependent enumerated subdesigns rather than independent experiments. The mixed-hardware screen is descriptive and does not prescribe a universal number of replicates.

Model-stack condition. Using the common trainable raw-identity universe, the InternVL3-8B condition compares ScalSelect with a composition-matched control. Its three paired contrasts are $+0.272$, $+1.262$, and -0.003 pp, with mean $+0.511$ pp and 95% interval $[-1.143, +2.164]$. This condition changes the model, processor, tokenizer, training implementation, evaluator adapter, and matching resolution together. It therefore bounds model-stack transport; it does not isolate a backbone effect or constitute a controlled sign comparison with Qwen3-VL.

Observation-by-budget condition. The completed same-A800 interaction estimate is $+0.229$ pp with 95% interval $[-1.191, +1.649]$ over five paired seeds ($p = 0.677$). None of the six benchmark interactions passes Benjamini–Hochberg correction at 0.05. This stress test neither supports a robust interaction nor establishes equivalence between the tested observation policies.

Candidate-pool scope check. We repeated the complete ScalSelect attribution grid on a second deterministic candidate-pool sample from the same frozen source reservoir. All 21 A800 cells passed the terminal-state, checkpoint, six-benchmark, and artifact-hash gates. Table 5 reports the

resulting contrasts. Under this Pool-2 contract, the within-envelope estimate is positive. Its descriptive difference from the Pool-1 estimate is consistent with contract sensitivity, but it cannot isolate a candidate-pool effect because the candidate sample and hardware contract both change. This is neither an external-dataset replication nor evidence to pool with Pool-1. The extension was specified after the primary results and is not preregistered.

Table 5: Same-reservoir Pool-2 scope check. Means and two-sided 95% Student-t intervals are in percentage points over 3 training-seed marginals after averaging 3 control draws within seed. Draw-by-seed cells are dependent and are not additional repetitions.

Contrast	Mean	Seeds	95% CI
$S - U$ (total policy)	$+0.835$	3	$[-0.278, +1.948]$
$M - U$ (realized envelope)	$+0.513$	3	$[-0.613, +1.639]$
$S - M$ (within-envelope ranking)	$+0.322$	3	$[+0.103, +0.541]$

Hardware sensitivity. Two exact-token A800–H20 same-cell comparisons differ by $+0.103$ and -0.184 pp in the all-task macro, but their maximum benchmark-level difference is 1.581 pp. They show that those two all-task comparisons would not change the claim tier; they do not establish hardware equivalence. H20, H20e, and A800 results are otherwise reported as separate contracts and never pooled.

G Artifact admission and provenance

Each scientific cell is linked to an immutable submission receipt and requires: the requested queue and four-by-eight resource shape; a terminal platform state; four rank ledgers ending in the declared completion stage; one unique expected checkpoint; benchmark result and per-example sample files with exact counts; non-empty metric objects; and SHA-256 records for code, subsets, checkpoints, and result files. Strict summaries fail closed if any condition is missing.

Infrastructure failures are not silently deleted. A failed run with an incomplete checkpoint is replaced under a fresh run identifier. If all training gates and the checkpoint pass but evaluation fails, an evaluation-only recovery points back to that checkpoint and does not create another training replicate. Cross-SKU pairs are marked in figures and ledgers. Smoke tests, queued-then-stopped jobs, and superseded attempts are excluded from scientific counts.

The primary figure manifest records SHA-256 hashes for every admitted source summary, generated CSV and LaTeX table, and color, grayscale, PDF, and SVG outputs. This binds the displayed claims to the exact validated evidence rather than to manually transcribed job logs.

H Additional limitations and negative evidence

The study does not estimate total end-to-end curation cost, since selector scoring and preprocessing overhead differ across policies. Equal example budgets do not imply equal compute. Our controls preserve selected factors but cannot rule out unmeasured attributes correlated with both membership and downstream performance. The completed observation-policy–budget interaction is retained as an appendix stress test rather than a main claim because its interval is wide and includes zero. Historical pilot selectors, superseded candidate pools, and incomplete hardware blocks are retained for provenance but are not used to support the selector-attribution conclusions. The legacy parser audit supports only the separate requirement to freeze evaluator behavior.