YOUR FILTER PICKS THE WINNER: PREDICTABLE SELECTION BIAS IN DIFFICULTY-FILTERED EVALUATION OF VISUAL TOKEN COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

A visual-token-compression benchmark scores methods only on items its "difficulty filter" (downsampling) fails. Phang et al. (2022) found adversarially filtered rankings unstable; we derive the distortion. In a threshold model, one-dimensional difficulty makes filtering the rescaling $(A_O - A_F)_+/(1 - A_F)$: rankings above the filter's accuracy survive, its 0% not unfair. With an operator-specific component it decreases strictly, at fixed accuracies, in the tetrachoric correlation of the two failure patterns: a filter measures dissimilarity from itself, invertibly given the marginals. On 1.1M generations from six VLMs, difficulty is not one-dimensional ($\hat{\kappa} = 0.51$–0.66; 0.78–0.88 net of cell-level noise on the open models), and that correlation spreads by 0.06–0.15, net of sampling noise, over a filter's competitors, which, by the theorem, can reorder them. The hard-subset winner departs from the unfiltered order in 11–17 of 45 cells per model against 6–8 by sampling, significantly on both LLaVA models (Qwen2-VL: $p = 0.11$). On VTC-Bench's model and filter, its conclusion that designed methods beat random selection holds at four of five budgets, filtered or not; a compressor filter can reverse it. The bias is choosable: a method's most flattering filter, picked on half the items, lifts its held-out rank by $+0.9$ to $+1.0$ (open models). It buys nothing: a matched-size random subset tracks the unfiltered ranking better on all six models ($\tau = 0.89$–0.97 against 0.62–0.81) and creates at most 1.5% unsupported orderings, inverting almost none, versus the failure set's 2–11%. Sample uniformly; for hard items, cross-fit a panel. Code, raw generations and analysis scripts will be released at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Visual token compression, dropping or merging most of the visual tokens a vision-language model (VLM) feeds its language backbone, has become a crowded field: dozens of methods report that 50–95% of the visual tokens can be removed at negligible cost (Chen et al., 2024a; Shang et al., 2025; Yang et al., 2025; Wen et al., 2025b; Alvar et al., 2025). Re-evaluations then found that random selection or plain spatial pooling matches or beats the designed methods on standard benchmarks (Wen et al., 2025a; Peng et al., 2025; Wang et al., 2026; Endo et al., 2025). The latest response is a change of *protocol*: VTC-Bench (Liao et al., 2026) keeps only the items a downsampled image fails on and reports that on this "hard" subset the designed methods are "not only effective but necessary". The filter row of its main table reads 0.0 on every benchmark, read there as an ideal reference point.

This paper is about what that protocol measures. We begin with the obvious objection, circularity, which is *wrong on its own*. If visual difficulty were one-dimensional, evaluating on the filter's failure set would be one monotone transformation of unfiltered accuracy, identical for every method (Theorem 1): it preserves the ranking of every method above A_F, ties everything below at zero, and awards the filter 0%, so its 0.0 row is correct, not a scandal. Circularity is unfair only when difficulty has a component depending on *which* information an operator destroys, an empirical question.

We formalise it with a two-component threshold model in which item i is answered correctly under operator O at budget b iff $s_O(b) \geq \theta_i + \delta_{i,O}$: θ_i is how much visual information the item needs, $\delta_{i,O}$ how badly O happens to destroy it. The share of variance carried by θ, which we call the

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.

one-dimensionality ratio κ, decides whether the protocol reorders: at $\kappa = 1$ it does not. Below 1 it need not: with a common correlation to the filter every competitor receives the same monotone rescaling, so the operative quantity is the *dispersion* of that correlation across competitors, which we also measure. Under this model the accuracy a filter F awards to a method O has a closed form in the marginal accuracies and the tetrachoric correlation ρ_{OF} of their failure patterns, and, at fixed marginals, it is strictly decreasing in ρ_{OF} (Theorem 2). So at fixed accuracies a filter penalises each competitor by its correlation with itself: methods sharing its mechanism are punished, methods that fail elsewhere rewarded, and two methods swap places when the less accurate one is sufficiently less correlated with the filter. Because ρ_{OF} is estimable from the *unfiltered* matrix, each filtered accuracy is predictable before it is run, which separates a computable bias from an observed instability.

We then measure: nine token selectors (six published rules, three baselines) at five budgets on LLaVA-1.5-7B, LLaVA-NeXT-Mistral-7B and Qwen2-VL-7B, nine input-level degradations at six budgets on three API models, on four benchmarks: 1,137,316 generations. The threshold structure is real and, on LLaVA-1.5, the aggregate curves align under a per-operator rescaling, yet the item-level failure sets disagree by enough to reorder a leaderboard: at a fixed filter the correlation ρ_{OF} spreads by 0.06–0.15 over the competitors it ranks, net of sampling noise, a spread that survives removing the cell-specific residual, which leaves $\kappa_\delta = 0.78$–0.88. Figure 1 shows the effect: rankings under different filters agree at only $\tau = 0.27$–0.28 although each is reproducible across item halves at 0.83–0.87, and the winner departs from the unfiltered order in 14, 17 and 11 of 45 cells against 6–8 by sampling. The Kendall τ between filtered and unfiltered rankings runs from $+1.00$ down to -0.37 and falls *below* a matched-size random subset's 5th percentile, with the filter's own entry at 0, in 27, 31 and 29 of 45 cells, by 0.19–0.24 on average. We also run the protocol's own filter, pixel-level downsampling, where it removes tokens, on Qwen2-VL (VTC-Bench's own model) and LLaVA-NeXT, and audit the conclusion such a subset licenses. Then we price the protocol: its bias is choosable (picking each method's filter from the unfiltered matrix lifts its rank by $+0.9$ to $+1.0$ positions on held-out items). And it buys no resolution: a uniformly random subset of the same size tracks the unfiltered ranking better on all six models and, on synthetic data with a known ordering and method-independent contamination, asserts nothing false where the failure set asserts 2–6% falsely.

Contributions. (i) A threshold model of visual sufficiency with a measurable one-dimensionality ratio κ, and a theorem that under one-dimensionality difficulty filtering rescales without reordering the methods it does not tie, so "circular" is not "unfair". (ii) A leakage theorem: at fixed accuracies a method's filtered accuracy is strictly decreasing in its tetrachoric correlation with the filter, with a closed form and an inversion condition. (iii) Estimators for κ, ρ and Λ validated on synthetic ground truth. (iv) The protocol's price: at matched size uniform random subsampling matches single-operator filtering on discrimination and beats it on agreement and artefacts, and the filter can be chosen to promote a method. (v) A study on six VLMs, eighteen operators and four benchmarks establishing $\kappa < 1$ everywhere and the filter-dependence of the hard-subset winner, an audit of the protocol on its own model, and XFILTER with a filter-stability diagnostic for benchmarks that want a hard subset. The contribution is a measurement framework and an audit, not new mathematics.

2 RELATED WORK

Adversarial and difficulty filtering. AFLite (Sakaguchi et al., 2020; Le Bras et al., 2020) filters a dataset against weak learners, and the filtered split becomes the benchmark; ImageNet-A (Hendrycks et al., 2021) keeps the natural images one classifier fails, which then scores near zero by construction (Bowman & Dahl, 2021); Phang et al. (2022) applied it to *evaluation* data with 18 adversaries and found the rankings unstable and disproportionately punishing to the adversary. Model-in-the-loop collection (Kiela et al., 2021; Shirali et al., 2023) is iterated construction, not one-shot filtering. Theorems 1–2 apply to any filter and competitor, budget-ordered or not; the budget axis adds within-operator nesting and the pooling that estimates κ. The mechanism is selection on a correlated variable, Thorndike's range restriction (Thorndike, 1949); ours is the correctness-table case and the fix.

Blind-answerable items. VQA benchmarks can often be answered without the image (Goyal et al., 2017; Agrawal et al., 2016). MMStar (Chen et al., 2024b), MMMU-Pro (Yue et al., 2025) and NaturalBench (Li et al., 2024) remove items text-only models solve; DatBench (DatologyAI et al., 2026), MMGist (Yuan et al., 2026) and MIRAGE (Asadi et al., 2026) audit dozens of benchmarks. These filter at the bottom of the budget axis ($b \to 0$) with a reference outside the ranked panel

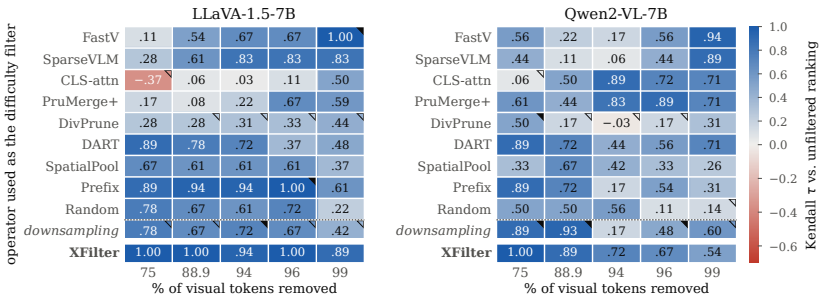

Figure 1: Which operator you choose as the difficulty filter decides the ranking. Kendall τ between the ranking of nine token compressors on the filter's failure set and their unfiltered ranking on the Step-A pool, for every filter (rows) and budget (columns), leading zeros dropped. Corner flags: filled where τ is at or above Theorem 1's prediction (Figure 8), open where it is undefined (the filter is the most accurate method); the rest fall below it, 38 of the 40 compressor-filter cells where it is defined on LLaVA-1.5 and 39 of 40 on Qwen2-VL; the prediction is a point value that ignores sampling, and §5.2 tests against a matched-size null. *Downsampling* (italic, below the dotted rule) is VTC-Bench's own filter, which removes tokens on Qwen2-VL but only information on LLaVA-1.5. Bottom row: the cross-fitted XFILTER protocol (§4).

(Appendix E); VTC-Bench moves the filter to a moderate budget and draws it from an operator family, where Theorem 2 applies. Long et al. (2026) and Kimhi et al. (2026) estimate per-item visual dependence for one operator; we use a panel to separate the shared and operator-specific components.

Token compression and re-evaluation. §5 lists the compressors we run; PyramidDrop (Xing et al., 2025) is a layer-wise schedule we omit. Random and pooling baselines match designed methods on standard benchmarks (Wen et al., 2025a; Peng et al., 2025); we do not adjudicate among them.

Curve collapse and evaluation statistics. Fitting accuracy curves across conditions on a probit scale is the effective-robustness idea (Taori et al., 2020; Miller et al., 2021), built on corruption benchmarks (Hendrycks & Dietterich, 2019); Usama et al. (2025) report a corruption×task interaction, an aggregate shadow of item-level heterogeneity. Rankings hinge on the benchmark and subset chosen (Dehghani et al., 2021), and a picked subset can keep a leaderboard's ranking (Rodriguez et al., 2021; Vivek et al., 2024); we ask what a failure-set filter does to one. Item-response models compress and audit LLM benchmarks (Polo et al., 2024; Zhou et al., 2026; Yauney et al., 2026); M1 is the deterministic (Guttman) limit of a Rasch model (Rasch, 1960; Guttman, 1944), M2 adds an operator-specific residual, and the tetrachoric correlation (Pearson, 1900) is its item-level statistic. Accuracy-adjusted agreement between failure patterns has a lineage: classifier diversity (Kuncheva & Whitaker, 2003), error consistency (Geirhos et al., 2020), chance-adjusted agreement of models' errors (Goel et al., 2025), correlated errors as a brake on test-set overuse (Mania et al., 2019); κ and Λ apply it to degradation operators, whose overlap has a predicted structure and effect on rankings.

3 A THRESHOLD MODEL OF VISUAL SUFFICIENCY

Setup. A benchmark has items $i = 1, \ldots, n$ (image, question, gold answer) and a frozen model M. A set $\mathcal{O}$ of *visual degradation operators* acts on the image or on M's visual tokens: input-level operators (downsampling, blur, JPEG, …) or token-level compressors keeping a subset of the tokens, each indexed by a *nominal budget* $b \in (0, 1]$, the fraction of visual information it claims to retain; $b = 1$ is the untouched image. $Y_i(O, b) \in \{0, 1\}$ is correctness on item i under O at budget b, and $A_O(b) = \frac{1}{n}\sum_i Y_i(O, b)$ is the accuracy of the *cell* (O, b).

The one-dimensional model. The simplest account of why compression hurts is that each item needs some amount of visual information and each cell delivers some amount:

$$\text{(M1)} \qquad Y_i(O, b) = \mathbf{1}\{\, s_O(b) \geq \theta_i \,\}, \tag{1}$$

with item *visual difficulty* $\theta_i \in \overline{\mathbb{R}}$ and operator *strength* $s_O(b)$ non-decreasing in b. Items with $\theta_i = -\infty$ are blind-answerable; items with $\theta_i = +\infty$ fail even at full budget. With $F(t) = \frac{1}{n}\#\{i : \theta_i \leq t\}$ the empirical distribution of difficulty, M1 implies

$$A_O(b) = F\big(s_O(b)\big), \tag{2}$$

so every operator traces the *same* curve F up to a monotone reparameterisation of the budget axis, and, far more restrictively, correct sets are nested, so at matched accuracy they are *identical*. The *excess disagreement* of two cells with $A_O(b) \le A_{O'}(b')$ is $D = \frac{1}{n}\#\{i : Y_i(O,b) = 1,\, Y_i(O',b') = 0\}$; under M1, $D \equiv 0$. Aggregate collapse equation 2 can hold while item-level nestedness fails badly.

The filtered protocol of Liao et al. (2026) has three steps, all in its released code. *Step A*: keep the items the model answers correctly at full budget, $\mathcal{P} = \{i : Y_i(\text{full}) = 1\}$; all accuracies below are on $\mathcal{P}$, so $A_{\text{full}} = 1$. *Step B*: choose a *filter* operator F at budget b_F and keep the items it gets *wrong*, $\mathcal{H}_F = \{i \in \mathcal{P} : Y_i(F, b_F) = 0\}$. *Step C*: report each competitor's accuracy on $\mathcal{H}_F$, $A_O^{\mathcal{H}_F}$.

Theorem 1 (monotone rescaling). *Under M1 with $A_F < 1$ ($\mathcal{H}_F \neq \varnothing$), for each cell (O, b),*

$$A_O^{\mathcal{H}_F} = \frac{(A_O - A_F)_+}{1 - A_F}, \qquad A_O = A_O(b),\ A_F = A_F(b_F). \tag{3}$$

The right-hand side of equation 3 is one map of A_O, independent of O: constant at 0 on $[0, A_F]$ and strictly increasing on $(A_F, 1]$. *Rank preservation*: filtering preserves the ranking of every operator with $A_O > A_F$. *Gap inflation*: differences between two such operators are multiplied by $1/(1 - A_F) \ge 1$, so filtering widens the gaps between methods with zero information gain. *Floor censoring*: every operator with $A_O \le A_F$ is mapped to 0 and tied, and the filter's own 0% is exactly what equation 3 prescribes. Unfairness therefore requires a departure from one-dimensionality.

The two-component model. Blur removes high spatial frequencies, patch dropping removes localised regions, attention-based pruning removes whatever one layer deems unimportant: operators at the same budget destroy *different* information, so we add an operator-specific term:

$$\text{(M2)}\quad Y_i(O,b) = \mathbf{1}\{\, s_O(b) \ge \theta_i + \delta_{i,O} + \varepsilon_{i,O,b} \,\}, \quad \mathbb{E}\delta = \mathbb{E}\varepsilon = 0,\ \ \mathrm{Cov}(\delta_{i,O}, \delta_{i,O'}) = \Sigma_{OO'}, \tag{4}$$

with δ uncorrelated with θ and ε independent of (θ, δ) and across cells. $\delta_{i,O}$ is how badly O happens to destroy the information item i needs; $\varepsilon_{i,O,b}$ is what is left over at one budget of one operator (§5.1 measures it). The *one-dimensionality ratio*

$$\kappa = \frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\theta) + \overline{\mathrm{Var}}(\delta) + \mathrm{Var}(\varepsilon)}, \qquad \overline{\mathrm{Var}}(\delta) = \tfrac{1}{|\mathcal{O}|}\textstyle\sum_O \Sigma_{OO}, \tag{5}$$

equals 1 (given $\mathrm{Var}\,\theta > 0$) exactly when M2 collapses to M1. Write ρ_{OF} for the correlation of the two latent shortfalls, standardised so that $A_O = \Phi(\tilde{s}_O)$ under the Gaussian law of Theorem 2; when the δ's are cross-family uncorrelated with equal variance, κ is the cross-family ρ, our estimator. All of this is on the Step-A pool $\mathcal{P}$; κ, ρ, Σ are pool quantities, and selection changes them (§5.1).

Theorem 2 (leakage). *Let the two latent shortfalls $(\theta + \delta_O + \varepsilon_O,\ \theta + \delta_F + \varepsilon_F)$ of equation 4 be jointly Gaussian on $\mathcal{P}$ with correlation ρ_{OF} (so ρ_{OF} is what the tetrachoric estimator targets), let $A_O, A_F \in (0,1)$, and let $\Phi_2(\cdot,\cdot;\rho)$ be the standard bivariate normal c.d.f. Then*

$$A_O^{\mathcal{H}_F} = \frac{\Phi(\tilde{s}_O) - \Phi_2(\tilde{s}_O, \tilde{s}_F; \rho_{OF})}{1 - \Phi(\tilde{s}_F)}. \tag{6}$$

Define the leakage index $\Lambda_{O|F} = A_O^{\mathcal{H}_F} - (A_O - A_F)_+ / (1 - A_F)$, *the excess of the filtered accuracy over its one-dimensional value. Then (i) $\Lambda_{O|F} \ge 0$; (ii) $\rho_{OF} = 1$ implies $\Lambda_{O|F} = 0$, recovering Theorem 1; and (iii) for fixed marginals A_O, A_F, $\Lambda_{O|F}$ is strictly decreasing in ρ_{OF} on $(-1, 1)$.*

Proposition 1 (no new information). *Let (Y_O, Y_F) be any 2×2 table on $\mathcal{P}$ with $A_O, A_F \in (0,1)$ and $\hat{\rho}_{OF}$ its tetrachoric correlation, the unique ρ with $\Phi_2(\Phi^{-1}(A_O), \Phi^{-1}(A_F); \rho) = \Pr[Y_O = 1, Y_F = 1]$. Then equation 6 holds with $\rho_{OF} = \hat{\rho}_{OF}$ as an algebraic identity.*

Given the two marginals, the filtered accuracy is an invertible re-encoding of the unfiltered table; the Gaussian model only *interprets* ρ_{OF} as a latent correlation. Proofs are in Appendix A.

Corollary 1 (no method is penalised more than the filter). *$\Lambda_{F|F} = 0$ while $\Lambda_{O|F} \ge 0$ for every O by Theorem 2(i): the filter attains the minimum leakage in the field, whatever the marginal accuracies.*

Corollary 2 (rank inversion). *For two operators with $A_{O_1} > A_{O_2} > A_F$ before filtering, filter F inverts their order iff*

$$\Lambda_{O_2|F} - \Lambda_{O_1|F} > \frac{A_{O_1} - A_{O_2}}{1 - A_F}. \tag{7}$$

The left side is strictly decreasing in ρ_{O_2F} and increasing in ρ_{O_1F} by Theorem 2(iii), so at fixed marginals inversion occurs below a threshold in ρ_{O_2F} (equivalently, above one in ρ_{O_1F}) when one exists. The cases $A_{O_2} \le A_F$ are in Appendix A.

The rest of the paper tests whether *a difficulty filter measures not quality but dissimilarity from itself*. Reading its 0% as the point from which competitors "demonstrate their true value" is backwards if they differ in that dissimilarity: $\kappa < 1$ is what makes room for the difference, not what creates it.

4 Estimation, prediction, and the cross-fitted filter

Estimating ρ and κ. For any two cells, $\hat{\rho}$ is the tetrachoric correlation of their 2×2 correctness table on $\mathcal{P}$ (Appendix B). $\hat{\kappa}$ is the mean tetrachoric correlation over all pairs of cells whose operators belong to *different mechanism families*, so shared within-family idiosyncrasies do not inflate it; same-operator and same-family pairs are reported separately because M2 with a family-block Σ predicts $\rho_{\text{same op.}} > \rho_{\text{same family}} > \rho_{\text{cross family}}$. Uncertainty is by paired item bootstrap throughout; all estimators were validated on synthetic data drawn from M2 with known θ, δ (Appendix C).

Prediction. By Proposition 1, plugging the in-sample $\hat{\rho}_{OF}$ into equation 6 reproduces $A_O^{\mathcal{H}_F}$ by identity and proves nothing; a prediction must estimate ρ elsewhere. *Held-out operator*: fit $\rho \approx c_0 + c_1 \mathbf{1}[\text{same family}] + c_2 \log b$ on all pairs not involving the operator under test, predict its ρ's, and predict its filtered accuracy under every filter. *Held-out items*: estimate $\hat{\rho}$ on one half of the items and predict filtered accuracies on the other. Both are compared with the one-dimensional null equation 3, the prediction a user who trusts the filter implicitly makes.

XFilter. Liao et al. (2026) note in their limitations that "only downsampling as the criterion" is a weakness and propose ensemble filtering over "a more diverse set of baseline methods". We agree, and add cross-fitting, which removes the evaluated method's own share of the panel rather than averaging it down (the panel ablation of Appendix B finds averaging does most of the work). The protocol conditions on $\theta_i + \delta_{i,F}$; we want to condition on θ_i. Choose a *panel* $\mathcal{Q} \subset \mathcal{O}$ excluding the operator under evaluation and its mechanism family, and score each item by its panel difficulty

$$\hat{d}_i^{(\mathcal{Q})} = \frac{1}{|\mathcal{Q}||B|} \sum_{O' \in \mathcal{Q}} \sum_{b \in B} \big(1 - Y_i(O', b)\big), \tag{8}$$

a monotone estimate of $\theta_i + \frac{1}{|\mathcal{Q}|}\sum_{O' \in \mathcal{Q}} \delta_{i,O'}$, and evaluate O on the hardest half by $\hat{d}$. Its residual correlation with the evaluated operator no longer depends on which operator was called the filter:

Proposition 2 (bias of XFilter)**.** *Under M2 with $O \notin \mathcal{Q}$, $\mathrm{Cov}(\delta_O, \delta_{O'}) = 0$ for $O' \in \mathcal{Q}$, the panel's idiosyncrasies mutually uncorrelated and ε absorbed into δ, $\rho_{O,\bar{\mathcal{Q}}} = \mathrm{Var}\theta \big/ \sqrt{(\mathrm{Var}\theta + \Sigma_{OO})(\mathrm{Var}\theta + \bar{\Sigma}_{\mathcal{Q}}/|\mathcal{Q}|)}$, which depends on O only through Σ_{OO}. Operators of equal idiosyncratic variance scored against one common panel therefore face one common panel correlation, so* XFilter *rescales them all by a single strictly increasing map of A_O and preserves their unfiltered ranking exactly, Theorem 1's rank preservation without M1 and without a floor.*

In practice neither hypothesis holds exactly: the panel changes with the evaluated operator, and the Σ_{OO} are not equal: estimated by each operator's own across-budget correlation (which also absorbs the cell-level residual ε, whose share varies by operator, Appendix B) they span 0.63–0.91 across the six arms. XFilter therefore preserves the ranking only approximately, and it ranks operators scored on slightly different halves (Appendix B). Where a nominal filter is named, as in the stability diagnostic, the panel also excludes its family. Two outputs accompany it: *threshold-stratified retention curves* $A_O(b \mid \hat{d}_i \in \text{bin}_k)$, the non-parametric content of "hard subsets" (Figure 3), and the *filter-stability diagnostic*, the mean pairwise Kendall τ between the rankings under each admissible filter choice, which a trustworthy filtered benchmark must keep near its no-reordering value.

5 Experiments

Design. Every arm is one binary matrix of items $\times$ cells. Items come equally from four benchmarks spanning the difficulty axis: MMStar (Chen et al., 2024b), ChartQA (Masry et al., 2022), POPE (Li et al., 2023) and TextVQA (Singh et al., 2019), three of the eight VTC-Bench uses. Budgets

are $b \in \{0.25, 0.111, 0.0625, 0.04, 0.01\}$, i.e. 75/88.9/93.75/96/99% reduction, VTC-Bench's own grid, plus $b = 1$ and four blind arms at $b = 0$. *Open-model arms*: LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-NeXT-Mistral-7B (Liu et al., 2024b; Jiang et al., 2023) and Qwen2-VL-7B (Wang et al., 2024), with nine token compressors in one harness, each keeping *exactly* $\max(1, \text{round}(bN))$ of the N visual tokens, in five mechanism families, three of which pair two published methods: FastV (layer-2 LLM attention; Chen et al., 2024a) and SparseVLM (question-token attention; Zhang et al., 2025); CLS-attention pruning (VisionZip's selection rule; Yang et al., 2025) and PruMerge+ (Shang et al., 2025); DivPrune (Alvar et al., 2025) and DART (Wen et al., 2025b); the position family, spatial pooling (the re-evaluations' pooling baseline) and prefix truncation (our control); and random selection. Family membership follows the selection rule, not the data; the two families we assigned ourselves are varied in Appendix B. Qwen2-VL, VTC-Bench's own primary model, is *dynamic-resolution* (N set by the image, 8 to 1,369, median 543), so image downsampling removes tokens in proportion to the budget (on LLaVA-NeXT, through its any-resolution tiling, about a third of them at every budget), and its filter runs as published on both, as a tenth *filter-only* operator, never ranked; LLaVA-1.5 carries it as a fixed-resolution control. *API arms*: GPT-4.1-mini and -nano (OpenAI, 2025, 1,000 items each) and Qwen3-VL-Plus (Bai et al., 2025, 500) with nine input-level operators in three families (resolution, masking, photometric) on a fixed canvas, plus a sixth budget $b = 0.0025$. Operators, prompts, graders and the Qwen2-VL harness are in Appendix D. Qwen3-VL-Plus rejects 0.6% of degraded images through a content filter; we analyse the 404 items with complete rows and treat the arm as supporting evidence (Appendix G).

5.1 The aggregate curves collapse; the item-level failure sets do not

An item's responses along the budget axis form a monotone $0 \cdots 01 \cdots 1$ pattern in 0.82–0.91 of cases per compressor on LLaVA-1.5, a mean excess of $+0.58$ over an independence null with matched marginals ($+0.51$ to $+0.62$ on the other five). The aggregate collapse equation 2 also fits: a shared probit slope in $\log b$ with per-operator offsets predicts a held-out compressor's curve with mean absolute error 0.070 on LLaVA-1.5, against 0.185 for a constant baseline (Appendix F).

Table 2 gives $\hat{\kappa}$ for every arm and Figure 6 the pairwise $\hat{\rho}$ on LLaVA-1.5. First, $\hat{\kappa}$ is far below 1: 0.592 [0.568, 0.610] on LLaVA-1.5, 0.661 [0.646, 0.676] on LLaVA-NeXT and 0.638 [0.627, 0.653] on Qwen2-VL with real compressors, and 0.513–0.528 on the three API models. Second, the nesting a family-block Σ predicts holds without exception, same operator at another budget (0.763 on LLaVA-1.5) > same family (0.670) > cross family (0.592); i.e. 59% of the latent variance is shared, 8% family-specific, 9% operator-specific and 24% specific to one cell (why equation 4 carries ε; guessing does not drive $\hat{\kappa}$ below 1, Appendix B). Third, the assumption-free statistic agrees: at matched accuracy cross-family cell pairs disagree on 10.3% of the pool (against $D = 0$ under M1), same-family pairs on 6.9% and same-operator pairs on 2.9%. Per benchmark (Table 9) ChartQA is the most one-dimensional on five of six arms and POPE the lowest on four, but neither ordering is universal, nor κ constant across architectures (Appendix H). So we claim the weaker statement: *visual difficulty is multi-dimensional in every model, vendor, architecture and operator level we measured.* One caveat cuts the other way: Step A truncates $\text{Var}(\theta)$ but not $\text{Var}(\delta)$, restricting our statistic's range. Without Step A $\hat{\kappa}$ is 0.70–0.83 on the open models, on items that also fail the blank-canvas anchor 0.42–0.55, without POPE within 0.07 of the pool value (Appendix B); every value is below 1, and the pool value governs the protocol, whose comparison happens inside $\mathcal{P}$. Part of its distance from 1 is the cell-specific ε: dividing it out ($\kappa_\delta = \hat{\kappa}/\hat{\rho}_{\text{same op.}}$) gives 0.78, 0.82 and 0.88; its share of a cell's latent variance, estimated per cell from one operator's budget pairs, differs across operators (0.09–0.36), but dividing each correlation by its own two cells' shares leaves the spread below in place (Table 4). What reorders a leaderboard is the spread of ρ_{OF} over competitors at a fixed filter, the standard deviation of $\hat{\rho}_{OF}$ across the eight competitors of a cell averaged over cells: 0.099, 0.083 and 0.060 on the open arms, 0.13, 0.13 and 0.19 on the API arms, or 0.097, 0.081, 0.058 and 0.126, 0.123, 0.148 net of the sampling noise in $\hat{\rho}$. Resampling images rather than items (POPE items share images) leaves every interval essentially unchanged, and refusals treated as missing move $\hat{\kappa}$ by at most 0.01 (Table 4). That spread has mechanism structure, testable because family membership comes from the published selection rules, not the table: with $\Delta = \bar{\Lambda}(\text{cross-family}) - \bar{\Lambda}(\text{same-family})$, $\Delta = +0.079$, $+0.016$ and $+0.082$ on the open arms and $+0.027$ and $+0.043$ on GPT-4.1-nano and Qwen3-VL-Plus, five of six arms significant (LLaVA-NeXT's small; Appendix H), and the sign survives both alternative family assignments (Table 14, Appendix B).

5.2 WHICH OPERATOR YOU FILTER WITH DECIDES THE RANKING

Figure 1 replicates the VTC-Bench protocol with each of the nine compressors as the filter, at all five budgets (all three arms in Figure 7; the 88.9% accuracies in Tables 11–13, winners in Table 10). Six of the nine compressors finish first somewhere on LLaVA-1.5, four on LLaVA-NeXT and five on Qwen2-VL. Part is mechanical: the protocol drops the filter from its own comparison, so when the pool leader is the filter the runner-up wins, and random subsets of the same sizes, ranked with the same method dropped, give 4.2, 3.4 and 5.4 distinct winners ($p = 0.01$, 0.38 and 0.99; 300 replicates). The sharper count is cells whose winner is not the best remaining unfiltered method, which exclusion alone would name: 14, 17 and 11 of 45, against 5.7, 7.8 and 8.1 under the same null ($p < 0.004$, < 0.004 and 0.11). Requiring the winner to beat the runner-up significantly (paired item bootstrap) leaves 26, 13 and 21 of the 45 cells with any significant winner; at 96% on LLaVA-1.5, filtering with DivPrune makes CLS-attn the significant winner and filtering with CLS-attn, DivPrune, a swap that exclusion alone also produces (1.7 expected under the null).

Kendall τ against the unfiltered ranking quantifies it. The filter's entry is 0 by construction, so Theorem 1 predicts imperfect agreement (Figure 8; Figure 1's corner flags mark the cells not below it), and as a point comparison the observed τ falls below it in 38, 40 and 39 of the 40 defined cells per model, by 0.26, 0.28 and 0.30. That prediction ignores sampling and the forced zero; the test is a sampling null, uniformly random pool subsets of size $|\mathcal{H}_F|$ with the filter's entry at 0 ($1{,}000$ draws per cell): mean $\tau = +0.72$, $+0.72$ and $+0.69$ against observed $+0.52$, $+0.48$ and $+0.48$, and the observed τ is below the 5th percentile in 27, 31 and 29 of 45 cells. Dropping the filter's row from both rankings (Figure 9) removes the mechanical part, leaving $+0.67$, $+0.62$ and $+0.62$ against a predicted $+0.85$. Nor is a filtered ranking itself noisy: at a fixed filter it is about as reproducible across random pool halves ($\tau = 0.87$, 0.87, 0.83) as the unfiltered ranking (0.90, 0.91, 0.79).

5.3 THE PROTOCOL AS PUBLISHED, ON THE ARCHITECTURE IT WAS PUBLISHED FOR

VTC-Bench filters with *image downsampling*, which removes no tokens on a fixed-resolution encoder but does on Qwen2-VL (proportionally) and LLaVA-NeXT (about a third of them at every budget), so it runs as published on both. As a filter it reorders about as much as an average compressor on Qwen2-VL and less on LLaVA-NeXT, each row's entry dropped from both rankings (Figure 9): on Qwen2-VL $\tau = +0.89, +0.93, +0.17, +0.48, +0.60$, a row mean of 0.61 against 0.62 for the nine compressor filters, fourth of ten, the top cell at 88.9%, above the one-dimensional prediction in three of its four defined cells (0.30 below at 93.75%); on LLaVA-NeXT $\tau = +0.83, +0.83, +0.78, +0.89, +0.60$, 0.79 against 0.62, also fourth of ten, above the prediction in two of five cells and below it by 0.12, 0.08 and 0.25 at 75, 88.9 and 99%, a mean shortfall of 0.08 against the compressor filters' 0.28. Its $\hat{\rho}$ to the compressors varies less than a compressor filter's on Qwen2-VL (SD 0.04 against 0.06 for compressor filters) and on LLaVA-NeXT it varies as much (0.09 against 0.08), highest with SpatialPool, Random and Prefix, so the spread alone does not rank filters; what matters is its alignment with the accuracy gaps (Corollary 2). The published *conclusion*, that on the hard subset designed methods are "not only effective but necessary", is more fragile. We measure it as the best designed method's accuracy minus random selection's, excluding cells whose filter is one of the two. On Qwen2-VL the gap spans -0.05 to $+0.15$ over the nine non-random filters (downsampling included) and five budgets, significantly positive in 30 of those 45 cells and significantly *negative* in one: at 99% with DivPrune as the filter, random selection beats the best designed method by 0.053 $[0.038, 0.068]$, still excluding zero after a Bonferroni correction across all 45 cells ($[0.029, 0.078]$; image-clustered: $[0.031, 0.074]$). Under VTC-Bench's filter the gap is $+0.08$, $+0.07$, $+0.05$, $+0.05$ and $+0.01$, significant at four of the five; on the unfiltered pool it is $+0.07$, $+0.06$, $+0.06$, $+0.05$ and 0.00, so at 99% the pool licenses no conclusion and the hard subsets license both. On LLaVA-1.5 it spans -0.04 to $+0.37$, two negative cells significant before correction and none after; on LLaVA-NeXT (45 cells, downsampling included) it is positive and significant in all 45 but ranges ten-fold. So on Qwen2-VL the sign of the licensed conclusion turns on the filter, and on the other two its size does.

5.4 FILTERED ACCURACIES ARE PREDICTABLE FROM UNFILTERED DATA; RANKINGS LESS SO

The one-dimensional null equation 3 misses filtered accuracies by 0.10–0.25 mean absolute error over six arms; a constant $\rho = \hat{\kappa}$ gives 0.044–0.085, the held-out-operator fit 0.028–0.056 and held-out items 0.009–0.037: resemblance is informative when estimated per pair, less so when predicted from

Table 1: What a hard subset costs, at matched size. Three evaluation sets of size $|\mathcal{H}_F|$ per (compressor filter, budget) cell: the filter's failure set, the all-operator panel's hardest $|\mathcal{H}_F|$ items, and a uniformly random subset. τ: agreement with the full pool's unfiltered ranking, the filter's row dropped from both sides. *created / inverted*: % of the 28 method pairs the subset orders significantly where the full pool does not, or oppositely; each set is tested by a paired item bootstrap of its own items, the random one within each drawn subset (five draws per cell).

	τ vs. unfiltered			created / inverted (%)		
model	protocol	ensemble	random	protocol	ensemble	random
real token compressors (open models)						
LLaVA-1.5-7B	+0.67	+0.87	**+0.93**	7.1 / 6.6	4.4 / 0.6	**0.7 / 0.0**
LLaVA-NeXT-7B	+0.62	+0.76	**+0.93**	4.0 / 9.6	1.7 / 2.5	**0.4 / 0.0**
Qwen2-VL-7B	+0.62	+0.63	**+0.89**	11.4 / 4.1	11.3 / 1.7	**1.5 / 0.0**
input-level operators (API models)						
GPT-4.1-mini	+0.79	+0.79	**+0.97**	2.2 / 1.5	1.2 / 0.3	**0.6 / 0.0**
GPT-4.1-nano	+0.75	+0.74	**+0.97**	2.6 / 1.9	1.7 / 1.3	**0.8 / 0.0**
Qwen3-VL-Plus	+0.81	+0.86	**+0.94**	2.8 / 0.3	0.8 / 0.1	**0.8 / 0.0**

family and budget. Accuracy is not order: per filter cell the held-out-operator prediction matches the observed ranking at $\tau = 0.64$, 0.62 and 0.60 on the open arms (0.80–0.83 on the API arms), against 0.58, 0.58 and 0.56 for the one-dimensional null and no better than a constant ρ; the observed ranking's split-half reproducibility (0.87, 0.87, 0.83) is the ceiling. The prediction is still enough to act on: estimating $\hat{\rho}$ and the marginals on half the items and giving each method the filter equation 6 predicts flatters it most improves its rank among the eight non-filter methods, against the *unfiltered* ranking of those eight on the held-out half, by +0.93, +1.02 and +0.93 positions on the open arms, every interval excluding zero, in 51–62% of cases against 4–18% worsened (API arms: +0.32, +0.15 and +0.39, each interval including zero). Since equation 6 with half-A estimates returns half A's own filtered accuracies (Proposition 1), a filter chosen on one half keeps flattering the method on the other: the filter is choosable in advance from unfiltered data, in the direction the theorem names.

5.5 Which evaluation set to use

Under the single-operator protocol, mean pairwise τ over the nine filters is +0.266 [0.248, 0.296], +0.271 [0.258, 0.292] and +0.284 [0.261, 0.302]: the ranking is mostly the filter's. Theorem 1 predicts about +0.77 on all three, each filter tying different methods at zero; without both filters' rows, +0.50, +0.51 and +0.54: about a third of the disagreement is that floor. Cross-fitting outside the evaluated operator's family (XFilter) lifts agreement to +0.909, +0.894 and +0.853, though its split-half reproducibility (0.87, 0.88, 0.73) exceeds the protocol's (0.87, 0.87, 0.83; Appendix B) by at most 0.01, trailing it on Qwen2-VL. The robust conclusion is not a method: a cross-fitted panel spanning several families is more filter-stable than any single operator, whichever panel.

5.6 What a hard subset costs

$\mathcal{H}_F$ is only 11–73% of $\mathcal{P}$, so a fair test needs a size control. We compare three sets of *the same* size at every cell: $\mathcal{H}_F$; the hardest $|\mathcal{H}_F|$ items by the all-operator panel of equation 8 (VTC-Bench's "ensemble filtering"); and a uniformly random subset. Table 1 defines *created* and *inverted*.

Table 1 is one-sided. A random subset of the same size tracks the unfiltered ranking better than the failure set on all six models and creates 0.4–1.5% and inverts almost none; the failure set creates 2–11% and inverts 0.3–9.6%. The ensemble panel is intermediate on artefacts (1–11% created, $\leq$ 2.5% inverted) and beats the protocol on agreement on the open arms and Qwen3-VL-Plus but ties or trails it on the two GPT arms, whose three families make a thin panel. Discrimination does not separate the rules (20.0–22.4 of 28 pairs resolved against 18.6–23.9 for random).

An unbiased subsample, though, wins by construction when scored against the unfiltered ranking, and VTC-Bench's premise is a contaminated pool. So we rerun the scorecard on M2 data with generating accuracies, hence true ordering, fixed in advance (Table 5). Against the *true* ordering, 2.0–6.0% of the orderings the failure set calls significant *contradict the truth* as κ falls from 0.8 to 0.4, none at $\kappa = 1$, and 0.00% for random at every κ: that is how κ governs the risk. Its τ column (+0.85 to +0.94) counts floor ties and the filter's own 0, present even at $\kappa = 1$ (+0.85 there, random +1.00), so the wrong-rate, not τ, tracks κ. Contamination, a share of the pool made *blind-answerable*, correct at a rate independent of the method, which a filter exists to remove, does

not rescue the protocol. At 40% contamination the failure set reaches only $+0.89$ and contradicts the truth on 3.2% of its significant orderings, against $+1.00$ and none for a random subset, which pays two discrimination points for the contamination it keeps (98% of pairs separated against 100% after screening with a *blind anchor*, a no-image reference outside the panel; the failure set separates 94%). Near ties (generating accuracies 0.01 apart, as close as the leading compressors on the real pools) drop the failure set to $+0.51$ with 27% of its significant orderings wrong, random to $+0.93$ with none. *Method-dependent* contamination, the blind items' success rate differing by method unrelated to quality, so that the pool itself agrees with the truth at only $+0.67$, is inherited by a random subset ($+0.71$) and the failure set alike ($+0.70$); only the anchor screen recovers the truth ($+1.00$). When the estimand is the ranking on the hardest θ quartile, which unequal Σ_{OO} separate from the marginal one, a random subset does not estimate it ($+0.50$), the failure set barely ($+0.62$) and the hardest quartile of a cross-fitted panel does ($+0.86$; Table 6). Random subsampling is thus the right rule for the marginal ordering, no cure for method-dependent contamination and no estimate of a hard-item ranking; anchor screening and cross-fitting are, and the single-operator filter is never best. On the real pools, though, the anchor's $\hat{\rho}$ spread is at least a compressor filter's, and screening on it reorders the ranking at extreme reduction on two of three models (Appendix B). What a hard subset was wanted for survives in the stratified curves (Figure 3): the stratum-conditional ranking is stable across leave-one-family-out panels in the three harder quartiles ($\tau = +0.79$ to $+0.96$); only the easiest, where every method sits within three points of ceiling through 96% reduction, is unstable, the noise the protocol was built to remove, displayed rather than filtered away.

6 Discussion

Recommendations. Theorem 1 defends the protocol at $\kappa = 1$ and Theorem 2 diagnoses it below; every arm has $\kappa < 1$ and a ρ_{OF} spread that realises the risk. Do not filter on a competitor's failures: the filtered table adds nothing to the unfiltered matrix (Proposition 1) and, §5.6 shows, subtracts not resolution, which a matched-size random subset keeps, but the guarantee that a significant ordering is true. For fewer items, sample uniformly; where contamination is the concern, screen with a blind anchor outside the ranked panel, then sample uniformly, and check the anchor's $\hat{\rho}$ spread like any filter's, since it reorders at extreme reduction (Appendix B); for *hard* items, cross-fit a panel rather than filter on one operator and report the stratified curves (Figure 3). Either way, report $\hat{\kappa}$, the ρ_{OF} spread and the filter-stability diagnostic.

Limitations. The three open models span two encoder designs and three backbones but are all 7B; LLaVA-1.5-13B on the same items replicates the main findings ($\hat{\kappa} = 0.55$, naive stability 0.28 against XFilter's 0.84, τ below Theorem 1's prediction in all 40 defined cells, though only two methods take the hard-subset wins; Appendix D, Table 8), and the API arms add vendors only at the input level. The compressors re-implement each selection rule at exact token counts, as comparing *rankings* requires, barring wall-clock claims, and are pure selectors; four published forms substituted on LLaVA-1.5 (FastV and DART pruned in the LLM, VisionZip and PruMerge+ merging) leave the one-dimensionality, stability and shortfall conclusions intact: $\hat{\kappa} = 0.58$, naive stability 0.26 against XFilter's 0.92, τ below that prediction in 38 of 40 cells (Table 7); the count of inverted cells varies, from none to three. We ran VTC-Bench's filter, budgets and its four compared methods' selection rules (FastV, VisionZip, PruMerge+, DART) on its primary model but our pool, three of its eight benchmarks: we do not show its numbers wrong, only that its output turns on a choice it leaves free. Two families are our labels, $\hat{\kappa}$ is range-restricted by Step A and anchor screening is scored against a known ordering only synthetically; on the real pools its ranking departs from the pool's at extreme reduction (Appendices B and H).

7 Conclusion

Difficulty-filtered evaluation of visual compression is safe when visual difficulty is one-dimensional, or when every competitor is equally correlated with the filter; neither holds. A filter scores each method by its dissimilarity from itself: the hard-subset winner departs from the unfiltered order in 11–17 of 45 cells per model, beyond sampling on two of the three; the conclusion such a subset licenses changes sign in one cell of 45, correction for the other 44 included; and a rank rises by choosing the filter in advance. None of it buys resolution: at matched size a random subset tracks the ranking better and rarely asserts an ordering the pool does not support.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates public benchmarks with public and commercial models under visual degradation; it involves no human subjects, personal data, or new data collection. Its risks are those of any evaluation paper: a protocol shown to be biased may have influenced published conclusions, and we have tried to state the scope of our claims precisely so that the correction is not overstated. Our critique concerns the VTC-Bench protocol (Liao et al., 2026), not the integrity of the authors who proposed it; their own limitations section already names single-operator filtering as a weakness and proposes ensemble filtering, which we evaluate.

REPRODUCIBILITY STATEMENT

Every quantity in this paper is computed by released code from released raw generations. The supplementary material contains: the binary correctness matrices for all six arms (1,137,316 records with model outputs, graded scores and refusal flags; the 184 Qwen3-VL-Plus requests the provider's filter rejected are absent, Appendix G), the item pool with benchmark identifiers, the evaluation harness for the three open models (token selectors with exact-count guarantees, attention capture, batching, and the Qwen2-VL self-check), the published-form runner of Appendix D with its 180,000 generations and the LLaVA-1.5-13B arm (276,000 generations), each as a separate file, the input-level operator implementations and seeds, the estimator library (bivariate normal c.d.f., tetrachoric solver, leakage index, partial correlations, bootstrap), the synthetic ground-truth checks of Appendix C, and the scripts that regenerate every table and figure. Appendix A contains full proofs; Appendix B the estimator details; Appendix D the compressor and operator definitions, prompts and graders; Appendix G the API-arm details including the content-filter issue and how it was handled.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of EMNLP*, 2016. arXiv:1606.07356.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.02175.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

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.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of NAACL-HLT*, 2021. arXiv:2104.02145.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a. arXiv:2403.06764.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2403.20330.

DatologyAI, Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Darren Teh, David Schwab, Fan Pan, Haakon Mongstad, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, Luke Merrick, Parth Doshi, Paul Burstein, Pratyush Maini, Scott Loftin, Spandan Das, Tony Jiang, Vineeth Dorna, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. DatBench: Discriminative, faithful, and efficient VLM evaluations. *arXiv preprint arXiv:2601.02316*, 2026.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2412.13180.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2006.16736.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. *arXiv preprint arXiv:2502.04313*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. arXiv:1612.00837.

Louis Guttman. A basis for scaling qualitative data. *American Sociological Review*, 9(2):139–150, 1944.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1903.12261.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:1907.07174.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL-HLT*, 2021. arXiv:2104.14337.

Moshe Kimhi, Nimrod Shabtay, Raja Giryes, Chaim Baskin, and Eli Schwartz. CARES: Context-aware resolution selector for VLMs. In *Proceedings of ACL*, 2026. arXiv:2510.19496.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. arXiv:2002.04108.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2410.14669.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*, 2023. arXiv:2305.10355.

Chenfei Liao, Wensong Wang, Zichen Wen, Xu Zheng, Yiyu Wang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Xin Zou, Yuqian Fu, Bin Ren, Linfeng Zhang, and Xuming Hu. Are we using the right benchmark: An evaluation framework for visual token compression methods. In *Proceedings of ACL*, 2026. arXiv:2510.07143.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024b.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of LVLMs by contrasting chain-of-embedding. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.23050.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.12580.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022. arXiv:2203.10244.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2107.04649.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025.

Karl Pearson. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A*, 195:1–47, 1900.

Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, Feng Xiao, and Lizhen Cui. Can visual input be compressed? A visual token compression benchmark for large multimodal models. *arXiv preprint arXiv:2511.02650*, 2025.

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. Adversarially constructed evaluation sets are more challenging, but may not be fair. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection (DADC), NAACL*, 2022. arXiv:2111.08181.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.14992.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Pædagogiske Institut, Copenhagen, 1960.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, 2021.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. arXiv:1907.10641.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2403.15388.

Ali Shirali, Rediet Abebe, and Moritz Hardt. A theory of dynamic benchmarks. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03165.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. arXiv:1904.08920.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2007.00644.

Robert L. Thorndike. *Personnel Selection: Test and Measurement Techniques*. Wiley, New York, 1949.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions. *arXiv preprint arXiv:2504.13690*, 2025.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yahong Wang, Juncheng Wu, Zhangkai Ni, Longzhen Yang, Yihang Liu, Chengmei Yang, Ying Wen, Lianghua He, Xianfeng Tang, Hui Liu, and Yuyin Zhou. When token pruning is worse than random: Understanding visual token information in VLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2512.07580.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of ACL*, 2025a. arXiv:2502.11501.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025b.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2410.17247.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.04467.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.08730.

Wenzhen Yuan, Jiacheng Ruan, Wutao Xiong, Chengping Zhao, Ting Liu, and Yuzhuo Fu. MMGist: A comprehensive multimodal benchmark for 2027. *arXiv preprint arXiv:2606.22437*, 2026.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of ACL*, 2025. arXiv:2409.02813.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2410.04417.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2505.15055.

A PROOFS

Theorem 1. Under M1 the events are nested: $\{Y_i(O,b)=1\}=\{\theta_i \le s_O(b)\}$. On the Step-A pool, $\Pr[\theta \le s_{\text{full}}]=1$ after normalisation. The filter set is $\mathcal{H}_F=\{s_F<\theta\le s_{\text{full}}\}$, with probability $1-F(s_F)=1-A_F$, and $\{\theta\le s_O\}\cap\mathcal{H}_F=\{s_F<\theta\le s_O\wedge s_{\text{full}}\}$, with probability $(F(s_O\wedge s_{\text{full}})-F(s_F))_+=(A_O-A_F)_+$ using $s_O\le s_{\text{full}}$. Dividing gives equation 3. The map $x\mapsto (x-A_F)_+/(1-A_F)$ is strictly increasing on $(A_F,1]$ and constant ($=0$) on $[0,A_F]$, which yields rank preservation above A_F, gap inflation by $1/(1-A_F)$, and floor censoring. □

Theorem 2. Let $Z_O=\theta+\delta_O+\varepsilon_O-s_O$ and $Z_F=\theta+\delta_F+\varepsilon_F-s_F$, whose standardised versions have correlation ρ_{OF}, and write $\tilde{s}_O=-\mathbb{E}Z_O/\mathrm{sd}(Z_O)$ so that $\Pr[Z_O\le 0]=\Phi(\tilde{s}_O)=A_O$. Then $A_O^{\mathcal{H}_F}=\Pr[Z_O\le 0, Z_F>0]/\Pr[Z_F>0]=(\Pr[Z_O\le 0]-\Pr[Z_O\le 0, Z_F\le 0])/(1-\Pr[Z_F\le 0])=(\Phi(\tilde{s}_O)-\Phi_2(\tilde{s}_O,\tilde{s}_F;\rho_{OF}))/(1-\Phi(\tilde{s}_F))$, which is equation 6. Differentiating Sheppard's representation (Appendix B) with respect to its upper limit gives the Plackett identity $\partial\Phi_2(u,v;\rho)/\partial\rho=\phi_2(u,v;\rho)$, which is strictly positive whenever u,v are finite and $\rho\in(-1,1)$; so Φ_2 is then strictly increasing in ρ. (For A_O or $A_F\in\{0,1\}$ a threshold is infinite, $\phi_2\equiv 0$, and Φ_2 does not depend on ρ at all, which is why Theorem 2 assumes interior marginals.) The numerator of equation 6 is therefore strictly decreasing in ρ_{OF}, the denominator does not involve ρ_{OF}, and since the one-dimensional value $(A_O-A_F)_+/(1-A_F)$ depends only on the marginals, $\Lambda_{O|F}$ is strictly decreasing in ρ_{OF} at fixed marginals: (iii). At $\rho=1$ the two standardised variables coincide, $\Phi_2(u,v;1)=\Phi(u\wedge v)$, and the numerator is $\Phi(\tilde{s}_O)-\Phi(\tilde{s}_O\wedge\tilde{s}_F)=(A_O-A_F)_+$: (ii). For (i) we do not need monotonicity or a limit: $\Pr[Z_O\le 0,Z_F\le 0]\le\min(A_O,A_F)$ is the Fréchet upper bound for any joint law, so $A_O^{\mathcal{H}_F}=(A_O-\Pr[Z_O\le 0,Z_F\le 0])/(1-A_F)\ge(A_O-A_F)_+/(1-A_F)$, i.e. $\Lambda_{O|F}\ge 0$ distribution-free. □

Proposition 1. Write $p_{11}=\Pr[Y_O=1,Y_F=1]$ and $p_{10}=\Pr[Y_O=1,Y_F=0]=A_O-p_{11}$. The filtered accuracy is $A_O^{\mathcal{H}_F}=p_{10}/(1-A_F)=(A_O-p_{11})/(1-A_F)$ by definition. The tetrachoric correlation is defined as the unique ρ with $\Phi_2(\Phi^{-1}(A_O),\Phi^{-1}(A_F);\rho)=p_{11}$; uniqueness follows from strict monotonicity of Φ_2 in ρ and its limits $\max(A_O+A_F-1,0)$ at $\rho=-1$ and $\min(A_O,A_F)$ at $\rho=1$, which are the Fréchet bounds on p_{11}. Substituting p_{11} into equation 6 gives an identity. □

Corollary 2. $A_{O_1}^{\mathcal{H}_F}<A_{O_2}^{\mathcal{H}_F}$ iff $(A_{O_1}-A_F)_+/(1-A_F)+\Lambda_{O_1|F}<(A_{O_2}-A_F)_+/(1-A_F)+\Lambda_{O_2|F}$. When both operators exceed A_F this is equation 7; when $A_{O_2}\le A_F<A_{O_1}$ the condition becomes $\Lambda_{O_2|F}-\Lambda_{O_1|F}>(A_{O_1}-A_F)/(1-A_F)$, and when both are at or below A_F the one-dimensional terms vanish and inversion happens iff $\Lambda_{O_2|F}>\Lambda_{O_1|F}$. In every case Theorem 2(iii) converts the condition into one on the correlations. □

Proposition 2. Write $\bar{\delta}_i=\frac{1}{|\mathcal{Q}|}\sum_{O'\in\mathcal{Q}}\delta_{i,O'}$ and let $\bar{\Sigma}_{\mathcal{Q}}=\frac{1}{|\mathcal{Q}|}\sum_{O'\in\mathcal{Q}}\Sigma_{O'O'}$ be the average panel *diagonal*. If the panel's idiosyncrasies are mutually uncorrelated then $\mathrm{Var}(\theta+\bar{\delta})=\mathrm{Var}\theta+\bar{\Sigma}_{\mathcal{Q}}/|\mathcal{Q}|$; with within-panel covariances the $1/|\mathcal{Q}|$ decay is lost and the variance tends to $\mathrm{Var}\theta$ plus the average off-diagonal covariance, so the panel should mix families rather than deepen one. Under the hypothesis $\mathrm{Cov}(\delta_O,\delta_{O'})=0$ for all $O'\in\mathcal{Q}$ we have $\mathrm{Cov}(\theta+\delta_O,\theta+\bar{\delta})=\mathrm{Var}\theta$; dividing gives the stated $\rho_{O,\bar{\mathcal{Q}}}$, which depends on O only through Σ_{OO}. If instead $|\sum_{O'\in\mathcal{Q}}\mathrm{Cov}(\delta_O,\delta_{O'})|\le c$, the covariance term is $\mathrm{Var}\theta+O(c/|\mathcal{Q}|)$.

The consequence for rankings is stronger than a cancellation. Operators with equal Σ_{OO} share one panel correlation $\bar{\rho}$, and by equation 6 their cross-fitted accuracies are then the *same* function of their marginals, $h(a)=\big(a-\Phi_2(\Phi^{-1}(a),\Phi^{-1}(\bar{A});\bar{\rho})\big)/(1-\bar{A})$ with $\bar{A}=\frac{1}{2}$ the fraction of the pool the panel does not select, whose derivative $h'(a)=\big[1-\Phi\big((\Phi^{-1}(\bar{A})-\bar{\rho}\,\Phi^{-1}(a))/\sqrt{1-\bar{\rho}^2}\big)\big]/(1-\bar{A})$ is strictly positive. So XFILTER applies one strictly increasing rescaling to every such operator and preserves their unfiltered ranking exactly, as Theorem 1 does above A_F under M1. The leakage does not cancel additively (at fixed $\bar{\rho}$ it is unimodal in A_O, peaking at $A_O=\bar{A}$), but a common monotone distortion is all a ranking requires.

The score equation 8 is $\hat{d}_i=\frac{1}{|\mathcal{Q}||B|}\sum_{O'}g_{O'}(\theta_i+\delta_{i,O'})$ with $g_{O'}(x)=\#\{b: s_{O'}(b)<x\}$ non-decreasing, so it is monotone in the *vector* of shortfalls. It reduces to a monotone function of the scalar $\theta_i+\bar{\delta}_i$ when the panel shares a strength grid on which g is affine over the range of the shortfalls and no item is censored at either end; otherwise the panel selects on a slightly different index, which biases the cross-fitted score towards operators whose grid is denser where the items lie. □

Table 2: One-dimensionality and filter stability across six models. $\hat{\kappa}$ is the mean cross-family tetrachoric correlation on the Step-A pool; "naive" and XFILTER are the mean pairwise Kendall τ between the rankings produced by the admissible filter choices under the VTC-Bench protocol and under cross-fitting (same-family filter pairs, which share a panel, included; see Appendix B). Brackets: 95% basic paired item bootstrap intervals (Appendix B).

model	items	pool	$\hat{\kappa}$ [95% CI]	naive τ	XFILTER τ	difference [95% CI]
real token compressors (open models)						
LLaVA-1.5-7B	6,000	2,723	0.592 [0.568, 0.610]	+0.266	+0.909	0.642 [0.605, 0.694]
LLaVA-NeXT-7B (Mistral)	6,000	3,614	0.661 [0.646, 0.676]	+0.271	+0.894	0.624 [0.588, 0.641]
Qwen2-VL-7B	6,000	4,368	0.638 [0.627, 0.653]	+0.284	+0.853	0.569 [0.538, 0.642]
input-level operators (API models)						
GPT-4.1-mini	1,000	660	0.528 [0.494, 0.561]	+0.378	+0.726	0.348 [0.173, 0.387]
GPT-4.1-nano	1,000	569	0.526 [0.483, 0.559]	+0.370	+0.852	0.482 [0.407, 0.563]
Qwen3-VL-Plus	404	319	0.513 [0.442, 0.570]	+0.440	+0.823	0.384 [0.327, 0.478]

B ESTIMATOR DETAILS

Bivariate normal c.d.f. We use Sheppard's representation $\Phi_2(h,k;\rho) = \Phi(h)\Phi(k) + \frac{1}{2\pi}\int_0^{\arcsin\rho} \exp\big(-\frac{h^2+k^2-2hk\sin t}{2\cos^2 t}\big)dt$ with 40-node Gauss–Legendre quadrature, and the closed forms $\Phi(h \wedge k)$ and $(\Phi(h)+\Phi(k)-1)_+$ at $\rho = \pm 1$ so that the degenerate self-cell (O,O) is handled exactly rather than by clipping.

Tetrachoric correlation. Given a 2×2 table with margins A_O, A_F and joint p_{11}, $\hat{\rho}$ solves $\Phi_2(\Phi^{-1}(A_O), \Phi^{-1}(A_F); \rho) = p_{11}$ by Brent's method on $[-1,1]$; tables at a Fréchet bound return ± 0.999 in the released code. Standard errors are by paired item bootstrap (resampling items, recomputing the table), which we found to be well calibrated where an ad-hoc delta-method formula was not.

κ. $\hat{\kappa}$ is the mean of $\hat{\rho}$ over all pairs of cells $(O,b),(O',b')$ with O and O' in different mechanism families (800 pairs for the nine compressors at five budgets; 828 for nine input operators at six budgets). Restricting to different families removes the shared idiosyncrasy that equation 4 with $\Sigma_{OO'} > 0$ would otherwise let inflate the estimate; the same-operator and same-family means are reported alongside. Under M2 with uncorrelated equal-variance idiosyncrasies, $\mathbb{E}\hat{\rho}_{\text{cross}} = \kappa$; note that ρ carries no budget index under equation 4, which is what licenses pooling cell pairs across budgets. Disattenuation divides by the test–retest tetrachoric reliability estimated from a replicate cell; on the one arm with a replicate run, GPT-4.1-mini at temperature 0 (Appendix G), it is 0.994 and the correction is immaterial; the open models decode greedily, and their replicates are bit-identical with the same batch composition (a different batch can flip a near-tied token; across sessions 97.8% and 97.6% of grades agree, Appendix D). **Why equation 4 carries ε.** A model with only θ and δ predicts $\rho = 1$ between two budgets of one operator; the measured value is 0.763 on LLaVA-1.5. Under greedy decoding a replicate with the same batch composition is bit-identical, so test–retest reliability cannot be the check here: the candidate explanation is not decoding noise but *chance-level correctness*, since a lucky yes/no or four-way guess that lands in one cell and misses in another is exactly a cell-specific residual. Recomputing $\hat{\kappa}$ on the two open-ended benchmarks alone (ChartQA and TextVQA, chance ≈ 0, where a correct answer cannot be a coin flip) gives 0.619, 0.621 and 0.521 against 0.592, 0.661 and 0.638 on the full pools: removing every guessable item leaves $\hat{\kappa}$ within 0.12, and below its full-pool value on LLaVA-NeXT and Qwen2-VL, so guessing is not what makes difficulty multi-dimensional.

Table 3: Robustness of $\hat{\kappa}$ to the pool definition, and of the ranking results to POPE, on the six arms. Columns 2–5: $\hat{\kappa}$ on the Step-A pool (items correct at full budget), on the Step-A items that also fail the blank-canvas anchor, on the Step-A pool without POPE, and on all items without Step A (item counts in parentheses). Last two columns: the filter-stability diagnostic (naive / XFILTER) recomputed without POPE.

arm	Step-A	Step-A, blank wrong	Step-A, no POPE	all items	naive$^{\text{no POPE}}$	XFILTER$^{\text{no POPE}}$
L-1.5	0.592 (2723)	0.417 (1663)	0.586 (1464)	0.826 (6000)	+0.307	+0.923
L-NeXT	0.661 (3614)	0.543 (2468)	0.628 (2294)	0.796 (6000)	+0.289	+0.896
Qwen2-VL	0.638 (4368)	0.551 (3241)	0.570 (3073)	0.699 (6000)	+0.303	+0.799
4.1-mini	0.528 (660)	0.461 (472)	0.489 (442)	0.647 (1000)	–	–
4.1-nano	0.526 (569)	0.518 (402)	0.503 (365)	0.664 (1000)	–	–
Qwen3-VL	0.513 (319)	0.404 (237)	0.485 (232)	0.547 (404)	–	–

Pool robustness (Table 3). Restricting the Step-A pool to items that also fail the blank-canvas anchor, items that demonstrably need the image, lowers $\hat{\kappa}$ on every arm (0.40–0.55), as further range restriction of θ should. Excluding POPE, whose yes/no format lets a fixed prior score at chance under any degradation, changes $\hat{\kappa}$ by at most 0.07 on any arm and leaves the filter-stability contrast intact (naive 0.31/0.29/0.30 against XFILTER 0.92/0.90/0.80 on the three open models). Without POPE the LLaVA-1.5 protocol at 88.9% gives $\tau = -0.33$ (CLS-attn) and -0.25 (PruMerge+) where the full pool gives $+0.06$ and $+0.08$, while on Qwen2-VL the two *LLM*-attention filters halve without changing sign ($+0.11$ and $+0.06$, from $+0.22$ and $+0.11$): dropping the benchmark moves individual cells by up to 0.4 without changing the picture. Split-half reliability of the rankings (mean Kendall τ between random halves of the pool, 50 splits, all budgets): unfiltered 0.90/0.91/0.79, single-operator filter 0.87/0.87/0.83, XFILTER 0.87/0.88/0.73 on LLaVA-1.5/LLaVA-NeXT/Qwen2-VL, and 0.91–0.96, 0.83–0.86, 0.82–0.87 on the three API arms.

Three biases deserve statement, and two of the three make the estimate conservative for our purpose. (i) When the Σ_{OO} differ across operators, $\rho_{OF} = \sqrt{\kappa_O \kappa_F}$ with $\kappa_O = \mathrm{Var}\theta/(\mathrm{Var}\theta + \Sigma_{OO} + \mathrm{Var}\varepsilon)$ and, since $x \mapsto (\mathrm{Var}\theta + \mathrm{Var}\varepsilon + x)^{-1/2}$ is convex, the mean of $\hat{\rho}$ over pairs *exceeds* κ by Jensen when every operator enters equally many pairs (family sizes are unequal here, so this holds approximately). (ii) Any residual positive cross-family covariance we have wrongly set to zero also inflates $\hat{\kappa}$, so (i) and (ii) bias the estimate upward and our claim $\kappa < 1$ is safe under both. (iii) In the other direction, $\hat{\kappa}$ is computed on the Step-A pool, where selection on difficulty has truncated $\mathrm{Var}(\theta)$: dropping Step A raises it from 0.592 to 0.826 on LLaVA-1.5, 0.661 to 0.796 on LLaVA-NeXT, 0.638 to 0.699 on Qwen2-VL and 0.51–0.53 to 0.55–0.66 on the API arms (Table 3). Finally, cells at the two lowest budgets have sparse tables, noted where they arise.

Sampling noise, resampling unit, refusals and ε per cell (Table 4). The spread of $\hat{\rho}_{OF}$ of §5.1 is inflated by estimation noise; subtracting the per-cell bootstrap bias of the variance gives 0.097, 0.081 and 0.058 on the open arms (raw 0.099, 0.083, 0.060) and 0.126, 0.123 and 0.148 on GPT-4.1-mini, GPT-4.1-nano and Qwen3-VL-Plus (raw 0.132, 0.131, 0.189): only the 319-item Qwen3-VL-Plus arm loses a sizeable share, and the headline range 0.06–0.19 becomes 0.06–0.15. Items share images (POPE asks up to six questions per COCO image; the LLaVA-1.5 pool's 2,723 items sit on 1,783 distinct images, its 1,259 POPE items on 480), so we also resample images within each benchmark: every interval is essentially unchanged (width ratio 0.93–1.21 for $\hat{\kappa}$, at most 0.007 in absolute width), Δ on GPT-4.1-mini stays non-significant ($[-0.008, 0.024]$), and the Qwen2-VL audit of §5.3 still has 30 of 45 cells significantly positive and one significantly negative, the sign-flip cell's Bonferroni interval being $[0.031, 0.074]$ with images resampled against the item-level $[0.029, 0.078]$; on LLaVA-1.5 two negative cells, the CLS-attn and DART filters at 99%, are significant before correction and neither is after. Treating refusals as missing rather than wrong moves $\hat{\kappa}$ by at most 0.010 (GPT-4.1-nano, 0.526 to 0.516; every other arm by at most 0.001) and leaves the spread unchanged. Finally, under equation 4 two budgets of one operator differ only in ε, so on the tetrachoric scale $\log \rho(O, b; O, b') = \frac{1}{2}[\log(1 - v_{Ob}) + \log(1 - v_{Ob'})]$ with v the ε share of a cell's latent variance, and least squares over budget pairs gives v per cell. The mean share per arm is 0.19–0.26, but per operator it ranges over 0.11–0.27 (LLaVA-1.5), 0.09–0.23 (LLaVA-NeXT), 0.19–0.33 (Qwen2-VL) and 0.14–0.36, 0.13–0.30 and 0.09–0.31 on the API arms, so ε is not one common factor across operators. Dividing every $\hat{\rho}$ by its own two cells' shares nonetheless leaves the spread in place, 0.095, 0.069 and 0.047 on the open arms and 0.133, 0.141 and 0.194 on the API arms, so the spread is not an ε artefact; the ε-free $\hat{\kappa}$ is 0.765, 0.816 and 0.866 on the open arms, consistent with $\kappa_\delta = 0.78$, 0.82 and 0.88, and 0.683, 0.681 and 0.627 on the API arms.

The blind anchor on the real pools. Table 5 scores anchor screening against a known ordering only synthetically. On the Step-A pools the blank-canvas anchor's $\hat{\rho}$ with the nine compressors spreads (SD across the nine, averaged over budgets) by 0.15, 0.10 and 0.06 on LLaVA-1.5, LLaVA-NeXT and Qwen2-VL, against 0.099, 0.083 and 0.060 for a compressor filter, so the anchor is not mechanism-neutral. Screening on it, keeping the items it gets wrong, retains 61, 68 and 74% of the pool and leaves the ranking of the nine compressors at $\tau \geq 0.89$ against the unfiltered ranking on LLaVA-1.5 at every budget and at ≥ 0.99 on LLaVA-NeXT through 96% reduction, but it falls to 0.71 on LLaVA-NeXT at 99% and to 0.78, 0.78 and 0.37 on Qwen2-VL at 93.75, 96 and 99% (a matched-size uniformly random subset's 5th percentile: 0.83 on LLaVA-NeXT, and 0.83, 0.78 and 0.83 on Qwen2-VL; 1,000 draws, `code/checks_filter_nulls.py`). An anchor is a filter, and should be checked with the same $\hat{\rho}$-spread diagnostic as any other.

Table 4: Sampling noise, resampling unit, refusals and the cell-specific residual, on the six arms. Each bracketed pair is the 95% interval from the paired item bootstrap used throughout (first) and from a bootstrap that resamples *images* within each benchmark (second), the cluster being the distinct image file; percentile intervals for $\hat{\kappa}$ and Δ, basic intervals for the stability difference (XFILTER minus naive); 200 replicates, 120 for the stability difference, drawn independently of the replicates behind Table 2 (120 there too) and Table 14 (300), so the item intervals differ from theirs in the second decimal by Monte Carlo error. *spread*: the SD of $\hat{\rho}_{OF}$ across the eight competitors of a cell, averaged over cells, raw and net of the per-cell bootstrap bias of the variance. *refusals missing*: $\hat{\kappa}$ with refusals treated as missing rather than wrong. ε *share*: the range over operators of the share of a cell's latent variance carried by ε, estimated from the across-budget correlations of one operator under equation 4; ε*-free*: $\hat{\kappa}$ and the spread after dividing each $\hat{\rho}$ by its own two cells' ε shares.

model	$\hat{\kappa}$ [item] [cluster]	spread raw / corrected	$\hat{\kappa}$, refusals missing	ε share per operator	ε-free $\hat{\kappa}$ / spread	Δ [item] [cluster]	stability difference [item] [cluster]
LLaVA-1.5-7B	0.592 [0.571, 0.609] [0.572, 0.607]	0.099 / 0.097	0.592	0.11–0.27	0.765 / 0.095	[0.072, 0.089] [0.072, 0.088]	[0.616, 0.688] [0.618, 0.690]
LLaVA-NeXT-7B	0.661 [0.649, 0.675] [0.645, 0.676]	0.083 / 0.081	0.661	0.09–0.23	0.816 / 0.069	[0.010, 0.026] [0.009, 0.025]	[0.583, 0.643] [0.586, 0.640]
Qwen2-VL-7B	0.638 [0.627, 0.650] [0.625, 0.650]	0.060 / 0.058	0.639	0.19–0.33	0.866 / 0.047	[0.077, 0.086] [0.077, 0.085]	[0.539, 0.649] [0.544, 0.630]
GPT-4.1-mini	0.528 [0.492, 0.567] [0.487, 0.569]	0.132 / 0.126	0.529	0.14–0.36	0.683 / 0.133	[-0.006, 0.020] [-0.008, 0.024]	[0.156, 0.395] [0.200, 0.393]
GPT-4.1-nano	0.526 [0.484, 0.571] [0.473, 0.565]	0.131 / 0.123	0.516	0.13–0.30	0.681 / 0.141	[0.012, 0.040] [0.012, 0.041]	[0.413, 0.626] [0.418, 0.539]
Qwen3-VL-Plus	0.513 [0.451, 0.583] [0.461, 0.593]	0.189 / 0.148	0.512	0.09–0.31	0.627 / 0.194	[0.014, 0.069] [0.011, 0.067]	[0.323, 0.462] [0.324, 0.455]

The family contrast in full. Theorem 2(iii) says that at fixed marginals the leakage a filter grants an operator decreases in their correlation. Evidence for this needs care: by Proposition 1, $\Lambda_{O|F}$ is an exact function of $(A_O, A_F, \hat{\rho}_{OF})$ from the same table, verified to a mean absolute error of 1×10^{-4} (the two tables per open arm at a Fréchet bound, where $\hat{\rho}$ is clipped to ± 0.999, deviate by up to 0.016), so regressing Λ on $\hat{\rho}$ with the marginals controlled tests an identity, not a fact about vision models (Figure 5 reports it as a consistency check only: 22 of 24 strata negative, all 22 significant). The contentful claim is that ρ has *mechanism structure*, that operators built on the same signal fail on the same items, and it is testable because family membership comes from the selection rules, not from the table. Quantified as $\Delta = \bar{\Lambda}(\text{cross-family}) - \bar{\Lambda}(\text{same-family})$ with a paired item bootstrap (Table 14), $\Delta = +0.079$, $+0.016$ and $+0.082$ on the three open arms and $+0.027$ and $+0.043$ on GPT-4.1-nano and Qwen3-VL-Plus: five of the six arms significant, GPT-4.1-mini not ($+0.007$ $[-0.006, +0.023]$); LLaVA-NeXT's is the smallest of the five and not uniform across its benchmarks (Appendix H). Two cautions. Λ still depends on the marginals, so Δ is coarse; the sharper statement is the nesting of §5.1. And only three families pair two published methods, the position family and random selection being our labels: resting the contrast on those three pairs gives $\Delta = +0.066$, $+0.015$ and $+0.099$, and moving random selection into the position family gives $+0.077$, $+0.015$ and $+0.053$, every interval excluding zero (Table 14). That, the filter-dependence of §5.2–§5.3, and the out-of-sample prediction of §5.4 are the evidence; the in-sample correlation is bookkeeping.

Leakage and the within-benchmark test. For each (filter, operator, budget) cell with $O \neq F$ we record $A_O^{\mathcal{H}_F}$, its one-dimensional value, $\Lambda_{O|F}$, $\hat{\rho}_{OF}$ and the marginals. Because Λ depends on the marginals as well as on ρ, the consistency check is the partial Spearman correlation of $\hat{\rho}$ and Λ given (A_O, A_F), computed within each model×benchmark stratum (the theorem concerns one filter facing one item population). We require at least 25 items in the filtered set for a cell to enter the test. Pooling strata attenuates the statistic toward zero because the level of Λ differs by benchmark; we report the stratified values.

Excess disagreement. For cell pairs whose accuracies on the pool differ by at most 0.03 and lie in $[0.10, 0.90]$, D is the fraction of pool items correct under the less accurate cell and wrong under the more accurate one; $D_{\max} = \min(A_O, 1 - A_{O'})$ is its maximum.

Rank statistics. Rankings are compared by Kendall's τ_b (Kendall, 1938; 1945) applied to the accuracy vectors directly, so that two methods scoring identically are treated as tied rather than being separated by an arbitrary tie-break. The filter-stability diagnostic is the mean pairwise τ over all pairs of admissible filter choices (those whose failure set has at least 25 items) at each budget, averaged over budgets. Under XFILTER the ranking for nominal filter F scores each operator O on the hardest half of the pool by a panel excluding both the family of O and the family of F; the single XFILTER

row of Figure 1 and Tables 11–13 names no filter and excludes only the family of O. The panel score takes only $|\mathcal{Q}||B| + 1$ distinct values (26–41 on the open arms, depending on panel size), so ties at the median make the realised selected fraction slightly above 0.5. Panel size matters and differs by arm: on the open-model arms, whose nine selectors span five families, the panel holds 5–8 operators for every (evaluated, nominal filter) pair, but on the API arms, whose nine operators span only three families, 30 of the 81 pairs leave a single-operator panel (colour quantisation), where XFILTER degenerates to another single-operator filter. The API XFILTER numbers should be read with that in mind, and a panel should be built to span families rather than to be large. Two nominal filters from the same family share a panel, so their XFILTER rankings coincide and the pair contributes $\tau = 1$ by construction: four of the 36 pairs on the open arms and 13 on the API arms (a resolution family of five and a masking family of three); dropping them moves the open arms to $+0.897$, $+0.881$ and $+0.835$ and the API arms from $+0.73$–0.85 to $+0.57$–0.77. Different-family filters still share panel families, so we also build the two rankings from disjoint halves of the families outside the evaluated operator's own, sharing no operator at all: $+0.82$, $+0.75$ and $+0.74$ on the three open arms, against the own-rows-dropped $+0.50$, $+0.51$ and $+0.54$ of the single-operator protocol. Trimming XFILTER's set to $|\mathcal{H}_F|$ for each nominal filter leaves $+0.83$, $+0.82$ and $+0.83$, so 12%, 12% and 4% of the gain is subset size and the rest is not. XFILTER agrees with the unfiltered ranking at $+1.00$ at 88.9% on LLaVA-1.5 and $+0.54$ to $+1.00$ overall. Table 2 and §5.5 include them; this paragraph reports the means without them, the disjoint-panel diagnostic, in which the two rankings are built from non-overlapping halves of the families outside the evaluated operator's own (six random halvings per budget), and the panel ablation, in which the leave-one-family-out panel is replaced by a leave-one-operator-out panel or by the full operator set including the evaluated operator (the ranking under the latter is the same for every nominal filter, so it has no stability diagnostic, only its agreement with the unfiltered ranking). **Intervals.** Item resampling perturbs a rank statistic asymmetrically (a resampled pool reorders methods that were nearly tied), so the bootstrap replicates of the stability difference centre away from the point estimate, in either direction (bootstrap means 0.64, 0.63 and 0.55 against point estimates of 0.64, 0.62 and 0.57 on the open arms), and a percentile interval inherits that shift. Table 2 and §5.5 therefore report basic (reverse-percentile) intervals, which correct the shift to first order; the percentile intervals and bootstrap means are in the released numbers file. **Panel ablation and split-half by budget.** Figure 4 shows the stability diagnostic per budget. Kendall τ against the unfiltered ranking, mean over budgets, for the all-operator, leave-one-operator-out and leave-one-family-out panels: on the open arms the three panels agree with the unfiltered ranking to within 0.1 of one another, and on the two GPT arms, whose resolution family holds five of nine operators, the leave-one-family-out panel is the best of the three. Split-half reproducibility of the XFILTER ranking is lowest at the smallest budgets on Qwen2-VL, where the hardest half sits near the floor and any ranking of it is noise; a single-operator failure set at those budgets covers a much larger share of the pool, which is why that protocol's reproducibility does not fall there.

C SYNTHETIC GROUND-TRUTH VALIDATION

Before any real data were analysed, the estimators were run on data generated from M2 with known $\theta_i \sim \mathcal{N}(0,1)$, $\delta_{i,O} \sim \mathcal{N}(0, \sigma_\delta^2)$ with optional within-family correlation, and operator strengths on a logarithmic budget grid, $n = 20{,}000$ items. Ten checks all pass: (S1) tetrachoric recovery on 45 cell pairs, mean absolute error 0.006, no $|z| > 3$ against bootstrap standard errors, boundary cells with true $\rho = 1$ recovered as ≥ 0.995; (S2) Theorem 1 exact at $\kappa = 1$ (MAE 0); (S3) with the in-sample $\hat{\rho}$, equation 6 reproduces filtered accuracies on 72 cells with MAE 0 (Proposition 1's identity, a code check) where the one-dimensional null has MAE 0.144, and the self-cell is exactly 0; (S4a) Λ decreases in ρ across operators at equal accuracy (Spearman -0.88, $p = 0.002$); (S4b) the analytic $\Lambda(\rho)$ is monotone on a grid; (S4c) the inverse map from filtered accuracy back to ρ is an identity to 5×10^{-6}; (S4d) the partial Spearman test recovers the sign when accuracies are unequal; (S5) $\hat{\kappa}$ recovers true $\kappa \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ to within 0.006; (S6) XFILTER raises the mean τ against the unfiltered ranking from 0.68 to 0.96 on a synthetic field with family structure; (S7) the monotone-response fraction is 1.0 for pure threshold data against an independence null of 0.59, and 0.72 against 0.48 with response noise. Tables 5 and 6 score the evaluation-set rules of §5.6 on the same generator, against the marginal ordering and against the ordering on the hardest quartile.

Table 5: The scorecard against a known ordering. Data drawn from M2 with the generating accuracies fixed in advance, so the true ranking of the nine operators is known and is external to any observed pool; $n = 6{,}000$, 24 replicates on common seeds, three mechanism families with within-family correlation 0.6. Columns: Kendall τ between each subset rule's ranking and the *true* ordering, and the percentage of the orderings the rule calls significant that contradict it; the τ column includes the floor ties a hard subset produces even at $\kappa = 1$ ($+0.85$ for the protocol and the ensemble there); every rule is tested by a paired bootstrap of its own evaluation set, and all rules use the same number of items at every cell. Standard errors over replicates are at most 0.011 (τ) and 0.36 points (%). Middle block: $\kappa = 0.6$ with a share of the pool made *blind-answerable*, answered correctly at a rate that does not depend on the method, which is the contamination filtering exists to remove; "dep.": the success rate on the blind items differs by method, unrelated to quality, so that the pool's own ordering agrees with the truth at only $\tau = +0.67$. *anchor*: a uniform sample of the items a no-image anchor gets wrong. Bottom: generating accuracies 0.01 apart ("ties"). At 40% contamination the protocol, ensemble, random subset and anchor separate 94, 99, 98 and 100% of pairs. "–": rule not applicable.

		τ vs. the true ordering				significant *and wrong* (%)			
κ	blind	protocol	ensemble	random	anchor	protocol	ensemble	random	anchor
0.4	–	+0.851	+0.993	+1.000	–	6.00	0.00	0.00	–
0.6	–	+0.896	+0.985	+1.000	–	3.95	0.10	0.00	–
0.8	–	+0.937	+0.972	+1.000	–	1.97	0.38	0.00	–
1.0	–	+0.850	+0.850	+1.000	–	0.00	0.00	0.00	–
0.6	20%	+0.895	+0.990	+1.000	+1.000	3.86	0.02	0.00	0.00
0.6	40%	+0.894	+0.996	+1.000	+1.000	3.19	0.00	0.00	0.00
0.6	30% dep.	+0.698	+0.898	+0.708	+0.999	12.42	2.07	12.16	0.02
0.6	ties	+0.514	+0.927	+0.930	–	27.25	0.09	0.00	–

Table 6: When the estimand is the hard-item ranking. Data from M2 at $\kappa = 0.6$ as in Table 5, but with unequal idiosyncratic variances (operator s.d. multipliers 0.5–1.5 under a fixed permutation), so that the true ranking on the hardest quartile of θ differs from the marginal ranking (the two truths agree at $\tau = +0.50$). Columns: Kendall τ between each rule's ranking and the marginal ordering, and between it and the hardest-quartile ordering. The cross-fitted row scores each operator on the top quartile of a panel outside its family; the last row is the ceiling, the ranking on the true hardest quartile itself.

rule	τ vs. marginal ordering	τ vs. hardest-quartile ordering
failure set	+0.638	+0.616
ensemble panel	+0.728	+0.773
random subset	+0.999	+0.500
cross-fitted panel, hardest quartile	+0.644	+0.856
true hardest quartile of θ	+0.500	+0.995

D IMPLEMENTATION DETAILS

Token compressors. The two LLaVA models are run in one harness with `transformers`; Qwen2-VL joins it below. For LLaVA-1.5-7B the CLIP ViT-L/14 encoder yields $N = 576$ visual tokens; for LLaVA-NeXT the any-resolution tiling yields a per-image N (674–2,928 in our pool), and every selector receives the actual N. The selection context contains the encoder's [CLS]-to-patch attention from its last layer (reading it from the penultimate layer, whose features LLaVA consumes, gives a selector agreeing with it at $\rho = 0.95$; see the published forms below), the projected visual features, and the language model's layer-2 attention to each visual token, from the last prompt token and from every text token after the image (captured with a forward hook on that single layer, verified identical to the full attention output to within 10^{-6}). Selectors: *FastV* keeps the top-k tokens by layer-2 attention from the last prompt token, following its reference implementation (the paper's text specifies attention averaged over all tokens); *CLS-attn* keeps the top-k by [CLS] attention (the VisionZip dominant-token rule, without the contextual merge, so that every method is a pure selector at an exact count); *PruMerge+* keeps the [CLS]-attention outliers of the interquartile rule topped up with uniformly spaced tokens, without its key-similarity merge; the outliers number 61 per image on LLaVA-1.5 (median; interquartile range 51–71), so at 93.75, 96 and 99% (36, 23 and 6 tokens) they outnumber the budget and the first k in raster order are kept, and PruMerge+ and CLS-attn keep the identical token set on 3.0% of items at 88.9% and at most 0.8% at the other budgets; *SparseVLM* promotes to "raters" the text tokens after the image whose total layer-2 attention to the image is at least

the mean, and keeps the visual tokens receiving the most attention from them, without the recycling stage; *DivPrune* greedily keeps the token whose minimum cosine distance to the already-kept set is largest (its Max-Min Diversity rule), seeded by the highest-[CLS]-attention token rather than by its own first-stage rule; *DART* keeps a fixed pivot set (the top $\mathrm{round}(0.02N)$ by encoder importance, 12 on LLaVA-1.5) and fills the budget with the tokens least similar to any pivot, its one-shot duplication-aware rule; at 99% the budget (6 tokens on LLaVA-1.5) is smaller than the pivot set, so the kept set is the top-k by encoder importance, CLS-attn's, and on every open arm the two 99% cells are one selector (identical generations on 99.8, 100 and 100% of items); removing DART's 99% cell moves $\hat{\kappa}$ by at most 0.002, the naive stability by at most 0.008 and the mean τ by at most 0.011 on the three open arms; *SpatialPool* keeps a uniform stride over the token grid (falling back to a uniform stride over the sequence when N is not a clean grid), trimmed to the exact count in raster order; *Prefix* keeps the first k tokens, which on LLaVA-NeXT, where the 576-token global view comes first in the sequence, are global-view tokens only on 62% of items at 75% and on all items at 88.9–99%, so Prefix is a different operator there than on LLaVA-1.5; *Random* keeps a uniformly random subset with a per-item seed, drawn afresh at each budget rather than nested across budgets; in the three main open-model arms that seed came from a per-process salted hash, so those draws (and the random-image control's) cannot be reproduced exactly, while the later runs seed from an MD5 of the item id. Every selector's output is passed through an exact-count guard that de-duplicates and tops up with the lowest-index unused tokens (raster order), so that all nine methods at a budget keep exactly $\max(1, \mathrm{round}(bN))$ tokens (on Qwen2-VL, 114 of the $6{,}000$ items at 99%, 14 at 96% and 1 at 93.75% have $\mathrm{round}(bN) = 0$ and keep one token instead of none; every record matches this count); the top-up supplies 5.9% of PruMerge+'s tokens at 75% and at most 0.9% at the other budgets on LLaVA-1.5; without this guard PruMerge+ would return fewer tokens than the budget, which would compare methods at different token counts. Kept tokens are spliced into the input embeddings in their original order and the prompt is decoded greedily for at most 32 new tokens. Compressed sequences at one budget have equal length for LLaVA-1.5 and are bucketed by length for LLaVA-NeXT, so no padding is ever introduced. On LLaVA-NeXT the language model receives the global view's 576 tokens and then the crop tiles re-arranged into one row-major grid, padding rows or columns removed and a separator token after each row; the [CLS]-attention scores are laid out in that order by passing them through the model's own packing function as a one-channel feature (applied to the projected features, the same call reproduces the model's image embeddings exactly). On that arm the four rules that read these scores (CLS-attn, PruMerge+, DART and DivPrune) and the downsampling filter were generated in a second session of the same harness on another machine; regenerating cells of ten LLaVA-1.5 and six LLaVA-NeXT items there reproduces the first session's grade on 97.8% and 97.6% of the deterministic cells (`code/checks_regeneration.py`). Blind arms replace the image with a blank canvas, Gaussian noise, or the image of a random other item; the "removed" arm omits the image token entirely.

Published forms. The panel's selectors omit the stages of four methods that prune inside the language model or merge tokens. On LLaVA-1.5-7B, the model all four were published on, we also run each in its published form, still at exactly $\max(1, \mathrm{round}(bN))$ tokens, and substitute it for its stand-in (`gpu/published.py`; $6{,}000$ items $\times$ six forms $\times$ five budgets, $180{,}000$ generations released as a separate file and not counted in the $1{,}137{,}316$). *FastV* prunes inside the LLM: every token passes through layers 0 and 1, the visual tokens with the least attention from the last prompt token at layer 1 (mean over heads) are dropped, and layers 2 onward and decoding run on the survivors, which keep their original position ids ($K = 2$). *DART* runs as released by its authors: pruned after layer 1 ($K = 2$); the pivots are the visual and text tokens with the largest L_1 norm of their layer-1 keys, four of each in the authors' configuration, scaled below their 64-token setting as $\min(4, \max(1, \mathrm{round}(4k/64)))$ so that the pivots never exhaust the budget, and for each pivot in turn the visual tokens least cosine-similar to it are kept. *VisionZip* adds its contextual merge: dominant tokens by the penultimate CLIP layer's [CLS] attention plus contextual tokens that merge the rest by key similarity, before the projector, in the authors' $54{:}10$ dominant-to-contextual ratio. *PruMerge+* adds its merging stage: the interquartile outliers of the [CLS] attention ranked by attention, topped up with evenly spaced tokens, each kept token updated with the attention-weighted sum of its 32 nearest neighbours in key space, plus one extra token carrying the attention-weighted sum of every discarded token, which counts toward the budget. Two corrections of the stand-ins themselves run alongside: DART's stand-in with a pivot count scaled to the budget, $\min(8, \max(1, \mathrm{round}(k/8)))$, and CLS-attn's stand-in read from the penultimate encoder layer, the layer whose features LLaVA consumes, instead of the last. Three checks precede the run: the projector reproduces the model's

own image embedding exactly, the manual in-LLM path with nothing pruned reproduces `generate` token for token, and every form keeps exactly $\max(1, \mathrm{round}(bN))$ tokens. Table 7 gives the result. Item by item, each published form agrees with its stand-in at a tetrachoric ρ of 0.93 (VisionZip), 0.85 (FastV), 0.79 (PruMerge+) and 0.72 (DART), with mean accuracy differences of $+0.002$, -0.009, $+0.015$ and $+0.049$; DART inside the LLM is far more accurate than its stand-in at 75% (0.89 against 0.72). Substituting all four at once gives $\hat{\kappa} = 0.578$ $[0.562, 0.597]$ against the panel's 0.592, a naive filter stability of $+0.255$ $[0.237, 0.278]$ against XFILTER's $+0.919$ $[0.890, 0.938]$, a mean τ of $+0.50$ against the unfiltered ranking, below Theorem 1's prediction in 38 of 40 cells by 0.29, six distinct winners and 16 cells whose winner is not the best remaining unfiltered method (the panel: 0.266, 0.909, $+0.52$, 38 of 40 by 0.26, six and 14). One at a time, $\hat{\kappa}$ is 0.583–0.599, the naive stability 0.213–0.306, XFILTER's 0.913–0.937, the mean τ $+0.48$ to $+0.55$, with 37–38 of 40 cells below the prediction and 12–18 surprise cells; the scaled-pivot DART stand-in gives 0.595, 0.311 and 0.928 and the penultimate-layer CLS-attn 0.596, 0.276 and 0.899 (item-level ρ with the stand-ins 0.80 and 0.95). The one-dimensionality, stability and shortfall conclusions hold under every substitution: $\hat{\kappa}$ stays far below 1, the naive filter stability stays 0.23–0.34 times XFILTER's, and the observed τ falls below the one-dimensional prediction in 37–38 of 40 cells. The count of inverted cells does not: with the stand-ins one cell is negative (CLS-attn as the filter at 75%, -0.37); with VisionZip in place of CLS-attn one stays negative (VisionZip as the filter at 75%, -0.11); with PruMerge+'s merge alone three are (CLS-attn at 75%, -0.54; PruMerge+ with its merge at 75%, -0.33; CLS-attn at 96%, -0.06); with all four published forms none is. The inversion is a property of particular filter-panel pairs, which is how Appendix H scopes it.

Table 7: The published forms of four compressors substituted for their stand-ins, LLaVA-1.5-7B (6,000 items, five budgets), one at a time and all four together, plus two corrections of the stand-ins themselves. ρ: item-level tetrachoric correlation with the stand-in, mean over budgets; Δacc: mean accuracy difference from the stand-in; $\hat{\kappa}$, naive and XFILTER τ as in Table 2; τ vs. unfiltered: mean over the 45 compressor-filter cells, the number of the 40 defined cells below Theorem 1's prediction, and the mean gap to the prediction ("shortfall") over all 40 defined cells; winners: distinct hard-subset winners of the nine, and cells of 45 whose winner is not the best remaining unfiltered method ("surprise").

	vs. stand-in			stability τ		τ vs. unfiltered			winners	
variant	ρ	Δacc	$\hat{\kappa}$	naive	XFILTER	mean	below Thm. 1	shortfall	distinct	surprise
nine stand-ins (main panel)	–	–	0.592	+0.266	+0.909	+0.52	38/40	0.26	6	14
FastV → pruned inside the LLM	0.85	−0.009	0.587	+0.264	+0.919	+0.51	38/40	0.27	6	17
CLS-attn → VisionZip (merge)	0.93	+0.002	0.599	+0.291	+0.919	+0.54	38/40	0.24	6	12
PruMerge+ → with its merge	0.79	+0.015	0.583	+0.213	+0.937	+0.48	38/40	0.30	6	18
DART → inside the LLM	0.72	+0.049	0.587	+0.306	+0.913	+0.55	37/40	0.24	4	14
all four published forms	–	–	0.578	+0.255	+0.919	+0.50	38/40	0.29	6	16
DART stand-in, pivots scaled to budget	0.80	−0.056	0.595	+0.311	+0.928	+0.56	37/40	0.22	5	13
CLS-attn stand-in, penultimate layer	0.95	+0.005	0.596	+0.276	+0.899	+0.53	38/40	0.26	5	19

Scale. The LLaVA-1.5 arm is rerun with LLaVA-1.5-13B (`llava-hf/llava-1.5-13b-hf`, the same CLIP encoder and LLaVA recipe on a 13B Vicuna backbone) in the same harness, on the same 6,000 items, with the same nine compressors at the same five budgets plus the unpruned cell that defines the Step-A pool: 46 cells and 276,000 generations, released as a separate file and not counted in the 1,137,316. The blind arms and the downsampling filter were not run at 13B. Table 8 sets the two scales side by side, each on its own Step-A pool (2,723 items at 7B and 2,829 at 13B; full-image accuracy 0.454 and 0.471 on all 6,000). The phenomena replicate. Difficulty is, if anything, less one-dimensional at 13B: $\hat{\kappa} = 0.550$ $[0.536, 0.576]$ against 0.592, the interval excluding the 7B point. The naive protocol is as unstable across filter choices ($+0.276$ against $+0.266$) and XFILTER again far more stable ($+0.841$ against $+0.909$), a difference of 0.566 $[0.480, 0.580]$. The mean τ against the unfiltered ranking over the 45 compressor-filter cells is $+0.53$, and the observed τ falls below Theorem 1's prediction in every one of the 40 defined cells (against 38 at 7B), by 0.26 on average at both scales. What changes is the number of distinct hard-subset winners, 2 at 13B (DivPrune in 27 cells, CLS-attn in 18) against 6, not the number of cells whose winner is not the best remaining unfiltered method, 14 of 45 at both scales: at 13B DivPrune is the most accurate compressor at 88.9% on the pool (0.840), with CLS-attn and PruMerge+ next (0.829 each), and the filter decides which of the two strong methods wins. The ranking inverts in four cells rather than one: CLS-attn as the filter at 75% (-0.22; the one negative cell at 7B, at -0.37), CLS-attn at 93.75% and 96%, and PruMerge+ at 88.9% (-0.06 each). The two models agree on the unfiltered ranking of the nine compressors at Kendall $\tau = 0.94$, 0.83, 0.89, 0.94 and 0.56 at 75, 88.9, 93.75, 96 and 99%, and item by item their

failure patterns in the same cell correlate at a tetrachoric 0.69 (mean over the 45 cells, on the 2,356 items in both pools).

Table 8: LLaVA-1.5 at 7B and 13B on the same 6,000 items, each on its own Step-A pool (the 7B row repeats Table 2). Brackets: 95% basic paired item bootstrap intervals. $\hat{\kappa}$, naive and XFILTER τ as in Table 2; τ vs. unfiltered: mean over the 45 compressor-filter cells, the number of the 40 defined cells below Theorem 1's prediction, and the mean gap to the prediction ("shortfall") over all 40 defined cells; winners: distinct hard-subset winners of the nine, and cells of 45 whose winner is not the best remaining unfiltered method ("surprise").

			stability τ			τ vs. unfiltered			winners	
model	pool	$\hat{\kappa}$ [95% CI]	naive	XFILTER	difference [95% CI]	mean	below Thm. 1	shortfall	distinct	surprise
LLaVA-1.5-7B	2,723	0.592 [0.568, 0.610]	+0.266	+0.909	0.642 [0.605, 0.694]	+0.52	38/40	0.26	6	14
LLaVA-1.5-13B	2,829	0.550 [0.536, 0.576]	+0.276	+0.841	0.566 [0.480, 0.580]	+0.53	40/40	0.26	2	14

Input-level operators. Images are normalised to a 512-pixel canvas. Resolution family: *downsample* (Lanczos to area fraction b, bicubic back to the canvas), *native downsample* (the small image is sent, so the token count actually falls), *pixelate* (box average with block side $1/\sqrt{b}$), *blur* (Gaussian, $\sigma = 1/(2\sqrt{b}) - 0.5$), *JPEG* (quality $95\,b^{0.7}$). Masking family: *patch drop* (random cells of a 32×32 grid set to grey, keeping fraction b), *block mask* (same on a 6×6 grid), *crop* (centre crop to area b, resized back). Photometric: *colour quantisation* (greyscale and posterisation to $1 + 7\sqrt{b}$ bits). The random masks are nested across budgets by seeding on (item, operator) only, so that the cells kept at a smaller budget are a subset of those kept at a larger one.

The dynamic-resolution arm. Qwen2-VL-7B-Instruct runs in the same harness with three architecture-specific pieces, each verified before the arm was launched (`gpu/qwen_selfcheck.py`). (i) Its vision transformer has no [CLS] token, so the encoder-attention family scores each patch by the attention it *receives* in the last vision block, the adaptation used when these rules are ported to Qwen. That block discards its attention weights, so we recompute them exactly from its captured input in query chunks, and average the 2×2 patch groups the model's merger fuses into one visual token. (ii) Position ids are three-dimensional (M-RoPE); dropping visual tokens must not renumber the survivors, or a pruned sequence would claim a different image geometry, so the kept tokens keep their columns of the full prompt's position ids. (iii) `generate` cannot be given three-dimensional positions together with input embeddings, so the harness runs its own batched greedy decoder; it is verified token-identical to `model.generate` on the unpruned prompt. Images for this arm are rebuilt on a 1024-pixel canvas instead of 512, because here the token count is set by the image, the property the arm exists to exercise, and it ranges from 8 to 1,369 merged visual tokens (median 543). That is also what makes image downsampling a genuine token-reducing filter on this arm, so VTC-Bench's own criterion can be run rather than approximated.

Prompts and grading. MMStar items are posed as four-option multiple choice with a single-letter answer; POPE as yes/no; ChartQA and TextVQA as short answers. Grading is exact match for MMStar and POPE (after option-letter extraction), relaxed numeric match with 5% tolerance for ChartQA, and standard VQA accuracy (min(1, matches/3) against ten references) for TextVQA, binarised at 0.5 for the threshold analysis; the continuous score is also released. Refusals are detected by a pattern list and logged; unparsable outputs count as wrong. Sampling uses temperature 0 throughout.

E BLIND ANCHORS

Table 15 compares four operationalisations of "no visual information" on the Step-A pool. On the two GPT models, removing the image is out-of-distribution: GPT-4.1-mini refuses 25% of requests and scores 11%, far below the extrapolation of its budget curve to $b \to 0$, while a blank or noise canvas scores 28% with almost no refusals; GPT-4.1-nano behaves the same way and additionally refuses some blank and noise canvases. A mismatched real image scores below the blank canvas on both (24.4 vs. 28.5; 27.1 vs. 29.3), because it supplies misleading rather than absent information. Qwen3-VL-Plus and the three open models refuse in about 1% of cases or fewer and place all four anchors within 6 points (3 on all but LLaVA-1.5). Image removal is therefore pathological only for API models that refuse; a blank canvas behaves consistently across all six models and is the anchor we recommend.

Table 9: Per-benchmark one-dimensionality and filter stability. Pool: items correct at full budget; acc: fraction of the benchmark's items in the pool; $\hat{\kappa}$: cross-family tetrachoric mean within the benchmark; naive and XFILTER: filter-stability diagnostic within the benchmark.

model	benchmark	pool	acc	$\hat{\kappa}$	naive τ	XFILTER τ
L-1.5	ChartQA	256	0.17	0.843	+0.307	+0.958
L-1.5	MMStar	494	0.33	0.565	+0.126	+0.739
L-1.5	POPE	1259	0.84	0.510	+0.261	+0.977
L-1.5	TextVQA	714	0.48	0.508	+0.361	+0.934
L-NeXT	ChartQA	785	0.52	0.645	+0.363	+0.855
L-NeXT	MMStar	549	0.37	0.620	+0.225	+0.878
L-NeXT	POPE	1320	0.88	0.476	+0.183	+0.953
L-NeXT	TextVQA	960	0.64	0.592	+0.357	+0.937
Qwen2-VL	ChartQA	1046	0.70	0.513	+0.438	+0.823
Qwen2-VL	MMStar	822	0.55	0.650	+0.259	+0.781
Qwen2-VL	POPE	1295	0.86	0.488	+0.314	+0.845
Qwen2-VL	TextVQA	1205	0.80	0.474	+0.358	+0.912
4.1-mini	ChartQA	137	0.55	0.681	+0.436	+0.655
4.1-mini	MMStar	134	0.54	0.347	+0.430	+0.781
4.1-mini	POPE	218	0.87	0.302	+0.265	+0.889
4.1-mini	TextVQA	171	0.68	0.454	+0.479	+0.956
4.1-nano	ChartQA	122	0.49	0.650	+0.414	+0.648
4.1-nano	MMStar	104	0.42	0.365	+0.415	+0.863
4.1-nano	POPE	204	0.82	0.336	+0.157	+0.805
4.1-nano	TextVQA	139	0.56	0.519	+0.488	+0.914
Qwen3-VL	ChartQA	102	0.85	0.686	+0.560	+0.883
Qwen3-VL	MMStar	61	0.60	0.472	+0.515	+0.757
Qwen3-VL	POPE	87	0.89	0.469	+0.359	+0.693
Qwen3-VL	TextVQA	69	0.81	0.499	+0.624	+0.913

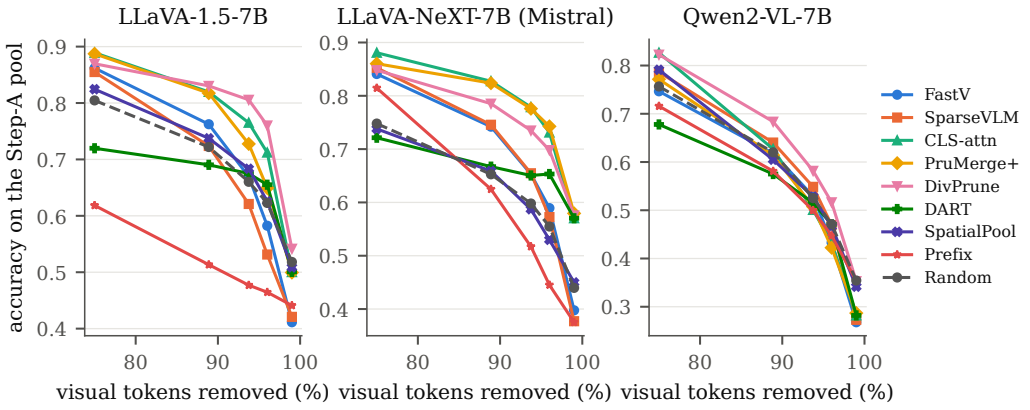

Figure 2: Accuracy on the Step-A pool against the fraction of visual tokens removed, for the nine compressors on the three open models. These are the curves from which a shared probit slope predicts a held-out compressor's curve with a mean absolute error of 0.070 on LLaVA-1.5 (§5.1); the item-level analysis of §5.1 shows how much they hide.

F AGGREGATE COLLAPSE

For each arm we fit $\Phi^{-1}(A_O(b)) = \alpha_O + \beta \log b$ with a shared slope β and per-operator offsets on the Step-A pool. On LLaVA-1.5 the shared slope is 0.30 probit units per unit of $\log b$ with a residual r.m.s. of 0.12 probit. Leaving out one compressor, fitting on the other eight and predicting the held-out curve with its offset estimated from its own $b = 0.25$ point gives mean absolute errors in accuracy of 0.019 (SpatialPool), 0.023 (Random), 0.032 (DivPrune), 0.049 (CLS-attn), 0.073 (PruMerge+), 0.082 (Prefix), 0.102 (FastV), 0.122 (SparseVLM) and 0.127 (DART), mean 0.070, against constant-baseline errors of 0.090–0.281 (mean 0.185); DART is the one compressor whose

Table 10: Which method finishes first on the "hard" subset, for every admissible filter and budget. The filter's own row is excluded from its own comparison. Reading down a column: the winner is a property of the filter, not of the field. †: the filter-only operator (image downsampling). Abbreviations: CLSat = CLS-attn, PruM+ = PruMerge+, DivPr = DivPrune, SprsVLM = SparseVLM, SpPool = SpatialPool, Rand = Random.

protocol	75%	88.9%	93.75%	96%	99%
LLaVA-1.5-7B (Step-A pool $n = 2{,}723$)					
unfiltered	CLSat	DivPr	DivPr	DivPr	DivPr
filter = FastV	DivPr	DivPr	DivPr	DivPr	DivPr
filter = SparseVLM	DivPr	DivPr	DivPr	DivPr	DivPr
filter = CLS-attn	SpPool	DivPr	DivPr	DivPr	Rand
filter = PruMerge+	DivPr	DivPr	DivPr	DivPr	DivPr
filter = DivPrune	PruM+	CLSat	CLSat	CLSat	Rand
filter = DART	CLSat	DivPr	DivPr	DivPr	Rand
filter = SpatialPool	CLSat	CLSat	CLSat	CLSat	DivPr
filter = Prefix	CLSat	DivPr	DivPr	DivPr	DivPr
filter = Random	CLSat	CLSat	CLSat	CLSat	DART
filter = downsampling†	CLSat	CLSat	CLSat	CLSat	CLSat
LLaVA-NeXT-7B (Mistral) (Step-A pool $n = 3{,}614$)					
unfiltered	CLSat	CLSat	CLSat	PruM+	PruM+
filter = FastV	CLSat	CLSat	PruM+	PruM+	PruM+
filter = SparseVLM	Prefix	CLSat	PruM+	PruM+	PruM+
filter = CLS-attn	Prefix	DivPr	DivPr	DivPr	PruM+
filter = PruMerge+	Prefix	DivPr	DivPr	DivPr	CLSat
filter = DivPrune	CLSat	CLSat	CLSat	PruM+	PruM+
filter = DART	CLSat	CLSat	PruM+	PruM+	PruM+
filter = SpatialPool	CLSat	CLSat	CLSat	PruM+	CLSat
filter = Prefix	CLSat	CLSat	CLSat	CLSat	CLSat
filter = Random	CLSat	CLSat	CLSat	PruM+	CLSat
filter = downsampling†	CLSat	CLSat	CLSat	CLSat	CLSat
Qwen2-VL-7B (Step-A pool $n = 4{,}368$)					
unfiltered	CLSat	DivPr	DivPr	DivPr	Rand
filter = FastV	CLSat	Rand	DivPr	DivPr	Rand
filter = SparseVLM	CLSat	Rand	Rand	DivPr	Rand
filter = CLS-attn	DivPr	DivPr	DivPr	DivPr	Rand
filter = PruMerge+	SprsVLM	DivPr	DivPr	DivPr	Rand
filter = DivPrune	CLSat	CLSat	SpPool	Rand	Rand
filter = DART	CLSat	DivPr	DivPr	DivPr	Rand
filter = SpatialPool	SprsVLM	DivPr	DivPr	DivPr	DivPr
filter = Prefix	CLSat	DivPr	DivPr	DivPr	DivPr
filter = Random	SprsVLM	DivPr	DivPr	DivPr	DivPr
filter = downsampling†	CLSat	DivPr	DivPr	DivPr	DivPr

Table 11: The VTC-Bench protocol with each compressor as the filter, LLaVA-1.5-7B, 88.9% **token reduction** (Step-A pool $n = 2{,}723$; accuracy in %, best per row in bold). Each row keeps the items the filter fails and scores everyone on them; the filter's own column is 0.0 by construction (Step B); † marks the filter-only operator (image downsampling, which on this fixed-resolution encoder removes information but no tokens); last column: Kendall τ against the unfiltered ranking. Two different methods top the table depending on the row, and the two encoder-attention methods, which read the same CLS-attention map, are each other's harshest filters.

protocol	FastV	SparseVLM	CLS-attn	PruMerge+	DivPrune	DART	SpatialPool	Prefix	Random	τ
unfiltered (Step A only)	76.2	72.4	82.0	81.7	**83.0**	69.0	73.7	51.3	72.2	+1.000
filter = FastV	0.0	8.7	48.4	48.4	**56.1**	31.2	45.6	26.9	41.4	+0.535
filter = SparseVLM	21.4	0.0	54.3	53.6	**60.9**	34.0	50.1	26.1	45.3	+0.611
filter = CLS-attn	32.0	29.9	0.0	13.6	**51.5**	31.6	44.0	27.5	43.2	+0.056
filter = PruMerge+	32.8	29.8	14.7	0.0	**51.5**	30.8	43.1	26.4	43.1	+0.085
filter = DivPrune	38.5	36.4	**48.5**	47.8	0.0	35.7	30.7	32.9	30.3	+0.278
filter = DART	47.2	41.2	60.1	59.2	**64.8**	0.0	49.2	21.5	47.4	+0.778
filter = SpatialPool	50.8	47.6	**61.5**	60.4	55.2	40.1	0.0	21.5	28.3	+0.611
filter = Prefix	64.3	58.0	73.1	72.4	**76.6**	50.0	57.7	0.0	54.7	+0.944
filter = Random	49.9	45.7	**63.1**	62.6	57.5	41.5	32.2	20.7	0.0	+0.667
filter = downsampling†	46.5	42.1	**62.8**	62.6	57.7	41.6	30.0	20.7	31.2	+0.667
XFilter (ours)	57.7	52.4	68.1	67.4	**71.6**	46.7	53.5	23.3	51.2	+1.000

own constant baseline is better, and it and SparseVLM are the worst fits. Figure 2 shows the curves. The point of §5.1 is not that collapse is perfect (the two LLM-attention rules and DART are the clearest exceptions) but that a reader of these curves would find one-dimensional difficulty a perfectly reasonable model, while the item-level statistics of §5.1 reject it decisively.

Table 12: The VTC-Bench protocol with each compressor as the filter on LLaVA-NeXT-Mistral-7B at 88.9% reduction (Step-A pool $n = 3{,}614$; accuracy in %). † marks the filter-only operator: VTC-Bench's own pixel-level downsampling, which on this any-resolution model removes tokens because a smaller image selects fewer crop tiles.

protocol	FastV	SparseVLM	CLS-attn	PruMerge+	DivPrune	DART	SpatialPool	Prefix	Random	τ
unfiltered (Step A only)	74.2	74.6	**82.7**	82.4	78.5	66.7	65.9	62.5	65.3	+1.000
filter = FastV	0.0	16.8	**52.3**	50.6	43.2	23.4	30.2	37.0	29.9	+0.278
filter = SparseVLM	15.7	0.0	**51.4**	49.2	42.5	22.5	28.9	38.3	28.2	+0.222
filter = CLS-attn	28.7	28.4	0.0	17.6	**32.2**	21.6	26.4	30.4	26.0	−0.278
filter = PruMerge+	27.8	26.7	19.3	0.0	**31.4**	19.5	23.7	29.8	23.9	−0.278
filter = DivPrune	31.8	32.0	**45.5**	43.7	0.0	28.6	21.0	29.5	21.4	+0.444
filter = DART	40.6	40.8	**59.4**	57.4	53.9	0.0	38.3	39.1	38.7	+0.667
filter = SpatialPool	47.2	47.0	**62.8**	60.6	50.3	39.8	0.0	31.5	23.6	+0.778
filter = Prefix	56.7	58.2	**68.0**	67.0	59.7	45.9	37.8	0.0	37.8	+0.986
filter = Random	48.0	47.4	**63.2**	61.4	51.4	41.3	24.9	32.7	0.0	+0.833
filter = downsampling†	46.5	45.9	**62.3**	60.7	48.6	41.2	24.5	29.6	24.3	+0.833
XFilter (ours)	54.0	54.1	**68.6**	67.5	61.6	43.0	40.1	40.4	38.5	+0.889

Table 13: The VTC-Bench protocol with each compressor as the filter on Qwen2-VL-7B at 88.9% reduction (accuracy in %). † marks the filter-only operator: VTC-Bench's own pixel-level downsampling, which on this dynamic-resolution model really does remove tokens. Three different methods top this table depending on the row.

protocol	FastV	SparseVLM	CLS-attn	PruMerge+	DivPrune	DART	SpatialPool	Prefix	Random	τ
unfiltered (Step A only)	61.8	64.0	62.7	61.4	**68.4**	57.5	60.5	58.1	62.0	+1.000
filter = FastV	0.0	13.0	37.4	35.4	38.6	27.4	36.2	25.8	**38.9**	+0.222
filter = SparseVLM	7.6	0.0	37.1	34.4	36.5	25.4	34.6	24.8	**37.5**	+0.111
filter = CLS-attn	35.9	39.4	0.0	24.3	**41.7**	29.7	33.4	30.5	36.0	+0.500
filter = PruMerge+	36.0	38.9	26.8	0.0	**41.3**	29.5	31.5	29.4	32.9	+0.444
filter = DivPrune	25.9	27.9	**31.2**	28.4	0.0	24.0	29.7	24.1	31.0	+0.167
filter = DART	34.8	36.9	38.3	35.9	**43.4**	0.0	36.7	31.6	38.1	+0.722
filter = SpatialPool	38.3	40.5	37.1	33.1	**43.7**	31.9	0.0	30.2	33.0	+0.667
filter = Prefix	32.3	35.4	38.1	35.0	**42.7**	30.7	34.2	0.0	36.8	+0.722
filter = Random	38.6	40.8	37.1	31.8	**42.5**	30.8	30.3	30.3	0.0	+0.500
filter = downsampling†	34.7	36.4	36.9	34.7	**43.2**	31.0	32.5	31.2	36.1	+0.930
XFilter (ours)	35.6	39.0	36.9	33.3	**44.0**	29.5	33.9	29.4	36.1	+0.889

G API ARMS

The three API models were queried through their APIs with identical prompts, images encoded as PNG, temperature 0, and a bounded worker pool with resumable logging keyed by (item, cell, model). A replicate run of four cells on 501 items of GPT-4.1-mini with the same prompts (2,004 item-cell pairs) gave a raw agreement of 0.965 and a tetrachoric reliability of 0.994.

Content filtering on Qwen3-VL-Plus. The provider's input-image safety filter persistently rejected 184 of 29,500 requests ("image data may contain inappropriate content"); retries did not clear them. The rejections were strongly operator-dependent: 157 in the masking family (dominated by patch drop, whose grey-cell images resemble noise), 23 in the resolution family, 2 photometric and 2 in the blind arms. Because our analysis requires a complete row per item, 96 of 500 items were dropped (40 TextVQA, 27 POPE, 24 MMStar, 5 ChartQA). The retained pool is selected on "accepted under every degradation", a selection that acts through the masking family; the arm's $\hat{\kappa} = 0.513$ still lands with the other models. A vendor-side filter has thus performed an operator-dependent selection on the evaluation pool, the mechanism this paper studies, and we treat the arm as supporting evidence rather than as a primary result.

Refusals. On the two GPT models, image removal triggers refusals in 25–29% of requests on the Step-A pool (Qwen3-VL-Plus: 1%), and GPT-4.1-nano also refuses 6% of blank and 14% of noise canvases. On non-blind cells the mean refusal rate is 0.4% for GPT-4.1-mini and Qwen3-VL-Plus (at most 3.4% in any cell, all at the two lowest budgets) and 5% for GPT-4.1-nano, whose rate reaches 18.5% at $b = 0.0025$. Refusals are graded as incorrect, the convention of the benchmarks; the refusal flag is released with every record so that other treatments can be evaluated, and treating refusals as missing moves $\hat{\kappa}$ by at most 0.01 on any arm (Appendix B, Table 4).

Table 14: Same-family versus cross-family leakage on every arm: mean $\Lambda_{O|F}$ over cells whose operator shares the filter's mechanism family and over cells that do not, and the paired-bootstrap difference Δ under the published family assignment; for the open arms also under the two alternatives of Appendix B, the three published pairs only (position family and random selection as singletons) and random selection moved into the position family. Theorem 2(iii) predicts $\Delta > 0$. Bottom: LLaVA-1.5-7B by benchmark.

arm	pool	$\bar{\Lambda}$ same	$\bar{\Lambda}$ cross	Δ [95% CI]	Δ, three pairs	Δ, Random in position
LLaVA-1.5-7B	2,723	0.180	0.260	$+0.079$ [$+0.072$, $+0.089$]	$+0.066$ [$+0.058$, $+0.078$]	$+0.077$ [$+0.069$, $+0.086$]
LLaVA-NeXT-7B	3,614	0.197	0.213	$+0.016$ [$+0.009$, $+0.025$]	$+0.015$ [$+0.006$, $+0.027$]	$+0.015$ [$+0.008$, $+0.024$]
Qwen2-VL-7B	4,368	0.159	0.241	$+0.082$ [$+0.076$, $+0.086$]	$+0.099$ [$+0.093$, $+0.103$]	$+0.053$ [$+0.048$, $+0.058$]
GPT-4.1-mini	660	0.147	0.154	$+0.007$ [-0.006, $+0.023$]	–	–
GPT-4.1-nano	569	0.130	0.157	$+0.027$ [$+0.011$, $+0.044$]	–	–
Qwen3-VL-Plus	319	0.105	0.148	$+0.043$ [$+0.009$, $+0.069$]	–	–
LLaVA-1.5-7B by benchmark						
ChartQA	256	0.111	0.144	$+0.033$ [$+0.019$, $+0.054$]	–	–
MMStar	494	0.170	0.276	$+0.107$ [$+0.089$, $+0.123$]	–	–
POPE	1259	0.242	0.326	$+0.084$ [$+0.070$, $+0.101$]	–	–
TextVQA	714	0.147	0.171	$+0.024$ [$+0.009$, $+0.040$]	–	–

Table 15: Blind anchors: accuracy on the Step-A pool in % (refusal rate in %) when the image is removed, replaced by a blank canvas, by Gaussian noise, or by the image of a random other item.

model	removed	blank	noise	mismatched image
LLaVA-1.5-7B	34.3 (0.1)	38.9 (0.1)	39.0 (0.1)	40.5 (0.0)
LLaVA-NeXT-7B (Mistral)	30.9 (0.2)	31.7 (0.0)	30.3 (0.0)	31.3 (0.0)
Qwen2-VL-7B	26.7 (0.4)	25.8 (0.0)	24.8 (0.0)	24.3 (0.0)
GPT-4.1-mini	11.4 (25.2)	28.5 (0.0)	28.2 (0.5)	24.4 (1.8)
GPT-4.1-nano	20.4 (29.0)	29.3 (6.2)	26.7 (14.1)	27.1 (3.5)
Qwen3-VL-Plus	24.5 (0.9)	25.7 (0.9)	25.7 (1.3)	26.3 (0.3)

H SCOPE OF THE CLAIMS

Six natural stronger statements are not supported by our data, and we record the boundaries. First, κ is not a universal constant: it ranges from 0.51 to 0.66 across the six arms on the Step-A pool (and 0.55–0.83 without Step A, Table 3), and the three open architectures span 0.59–0.66; the claim is that κ is below 1 everywhere ($\kappa_\delta = 0.78$–0.88 net of the cell-level residual on the open models), not that it takes a particular value. Second, across benchmarks $\hat{\kappa}$ puts ChartQA highest on five of the six arms and POPE lowest on four, but MMStar is highest on Qwen2-VL and TextVQA lowest on LLaVA-1.5 and Qwen2-VL, so we claim a real spread across benchmarks and not an order. Third, the same-family versus cross-family contrast of Table 14 is significant on five of the six arms but not on GPT-4.1-mini, where the family labels are coarse and the resolution family happens to hold the more accurate operators, which confounds an uncontrolled contrast; nor is it uniform where it holds: on LLaVA-NeXT it is small ($+0.016$) and varies by benchmark, $+0.109$ on POPE and $+0.076$ on MMStar against $+0.006$ on ChartQA (not significant) and -0.033 [-0.044, -0.018] on TextVQA, where same-family pairs leak *more*. Fourth, and most important, the accuracy-controlled partial correlation of Figure 5 is *not* evidence for Theorem 2(iii): Proposition 1 makes it an identity, so it can only ever confirm the arithmetic (Appendix B), and we do not count its 22-of-24 tally as evidence. The evidence is the family contrast, whose labels come from the published selection rules; the filter-dependence of the winner in §5.2; the audit of the licensed conclusion in §5.3; and the out-of-sample prediction of §5.4, where the decomposition is that most of the gain over the one-dimensional null comes from $\kappa < 1$ and not from operator-resolved ρ. A fifth boundary is the strength of inversion: one of the 45 cells is negative on LLaVA-1.5 (-0.37, CLS-attn as the filter at 75%) and one on Qwen2-VL (-0.03), three on LLaVA-NeXT (-0.28 each: CLS-attn as the filter at 88.9 and 93.75%, PruMerge+ at 88.9%), so "the winner depends on the filter" is general while "the ranking inverts" is a statement about particular filter-panel pairs: clear inversions occur on both LLaVA arms, on LLaVA-NeXT with the two encoder-attention methods as filters, while Qwen2-VL's one negative cell is -0.03. The LLaVA-1.5 inversion also depends on the selectors' form: it is a cell which substituting all four published forms removes (no negative cell) while PruMerge+'s merge alone gives three (Appendix D); LLaVA-NeXT's three arise under the stand-in rules alone, the published forms having been run on LLaVA-1.5 only. At 13B four cells are negative, the most negative being the same CLS-attn cell at 75% (-0.22; Table 8). Sixth, the count of distinct argmax

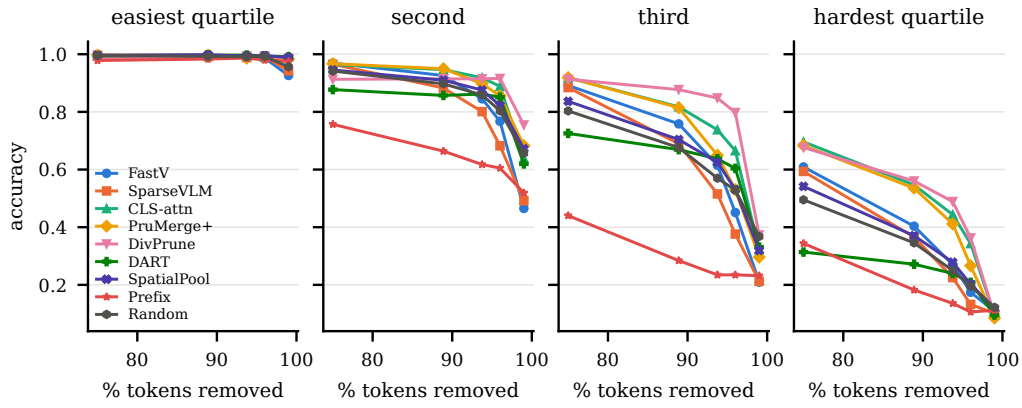

Figure 3: Threshold-stratified retention curves, the non-parametric output of §4, on LLaVA-1.5-7B: accuracy against the fraction of tokens removed within each quartile of cross-fitted panel difficulty $\hat{d}$ (each compressor scored by a panel excluding its own family). The easiest quartile is flat for every method; the separation a hard subset reports comes from the two hardest quartiles, where the two LLM-attention rules fall fastest and prefix truncation is lowest throughout, while CLS-attn, PruMerge+ and DivPrune hold 0.41–0.49 at 93.75% reduction on items where the rest hold 0.14–0.28; at 99% every method is within 0.09–0.12 of the floor.

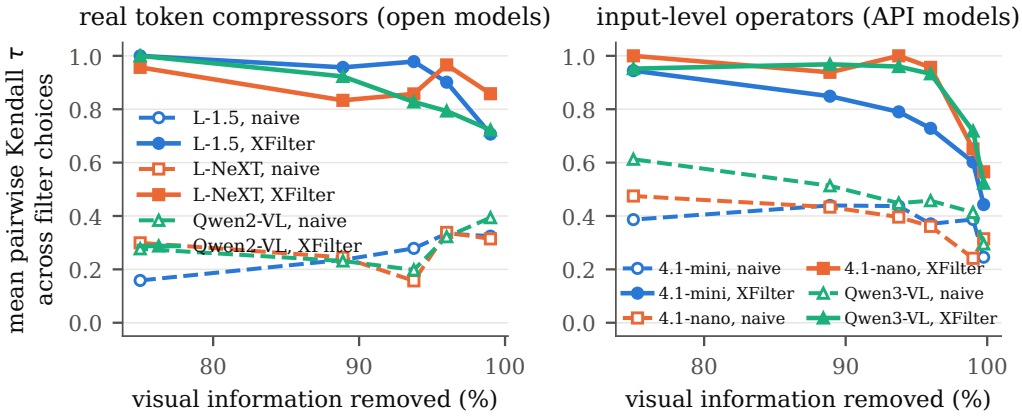

Figure 4: A ranking should not depend on which operator built the subset. Mean pairwise Kendall τ between the operator rankings produced by different filter choices, per budget, under the single-operator protocol (dashed) and under XFILTER (solid), for the three open-model arms with real token compressors (left) and the three API arms with input-level operators (right). Instability is present at every budget, including the moderate ones where filtered benchmarks are actually used.

winners (6, 4 and 5 of nine) is largely the protocol's exclusion of the filter at work: a null that drops the same method from a random subset of the same size expects 4.2, 3.4 and 5.4 ($p = 0.01$, 0.38 and 0.99), and the mutual swaps of §5.2 arise under it too (1.7, 2.0 and 0.8 expected against 1, 2 and 0 observed; on LLaVA-NeXT, DivPrune with PruMerge+ at 96% and CLS-attn with Prefix at 75%). What exceeds that null is the number of cells whose winner is not the best remaining unfiltered method, 14, 17 and 11 of 45 against 5.7, 7.8 and 8.1, on the two LLaVA arms ($p < 0.004$) but not on Qwen2-VL ($p = 0.11$); at 13B the two counts come apart, two distinct winners but the same 14 of 45 departing cells. The evidence for filter-dependence is therefore the systematic shortfall of τ below a matched-size sampling null (§5.2), and the split-half reproducibility of each fixed filter's ranking. Nor is random subsampling a cure for method-dependent contamination, where it inherits the pool's error (Table 5), which is why the recommendation screens with an anchor where contamination is the concern. One further boundary of the design: Qwen3-VL-Plus is vendor-filtered (Appendix G).

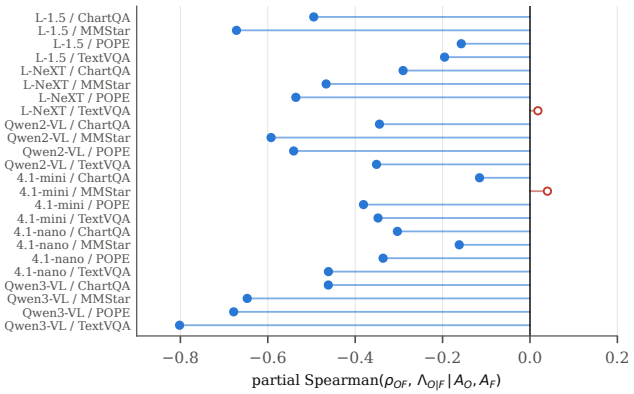

Figure 5: Consistency check, not a test (Appendix B): within each model×benchmark stratum, the partial Spearman correlation between $\hat{\rho}_{OF}$ and the leakage $\Lambda_{O|F}$ controlling for both marginal accuracies, over all (filter, operator, budget) cells with $O \neq F$. Proposition 1 makes the sign an identity; the spread reflects imperfect rank-linear control of A_O, which enters Λ unimodally. Filled markers are significant at $p < 0.05$. All twenty-four model×benchmark strata are usable.

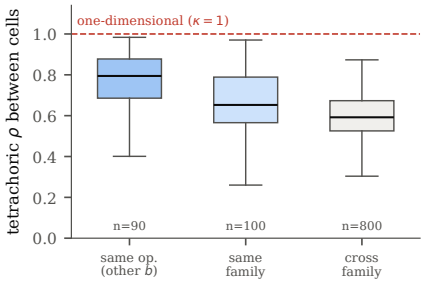

Figure 6: Tetrachoric ρ between all pairs of cells on LLaVA-1.5 (Step-A pool, n=2,723), by relation: the same operator at two budgets, two operators of one mechanism family, and two operators of different families. Under one-dimensional difficulty every box would sit at $\rho = 1$; the means are the $0.763 > 0.670 > 0.592$ of §5.1.

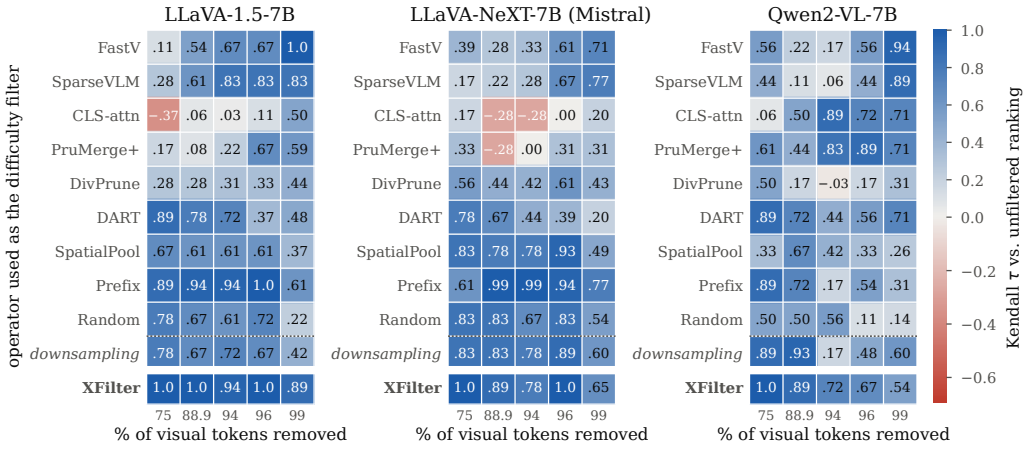

Figure 7: Figure 1 for all three open models, without the corner flags; Theorem 1's predictions are in Figure 8. Leading zeros are dropped, τ lying in $[-1, 1]$. The mean observed τ over the 45 compressor-filter×budget cells is $+0.52$, $+0.48$ and $+0.48$; one cell is negative on LLaVA-1.5 and on Qwen2-VL, three on LLaVA-NeXT.

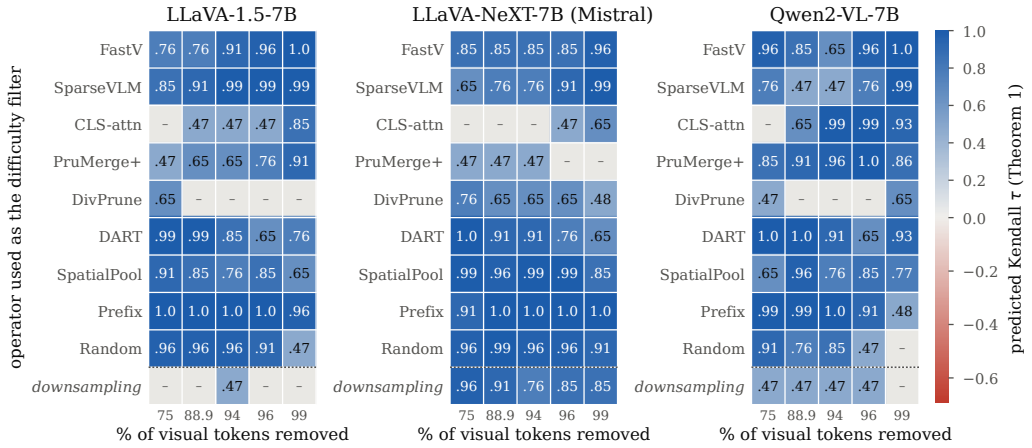

Figure 8: Theorem 1's prediction for every compressor-filter and downsampling cell of Figure 7, computed from the unfiltered accuracies alone through equation 3, each competitor at $(A_O - A_F)_+ / (1 - A_F)$. A grey "–" marks an undefined cell: the filter is the most accurate method, so every competitor is predicted at 0. The mean over the 40 defined compressor-filter cells is $+0.82$ on all three models, against the observed $+0.56$, $+0.54$ and $+0.53$ over the same cells.

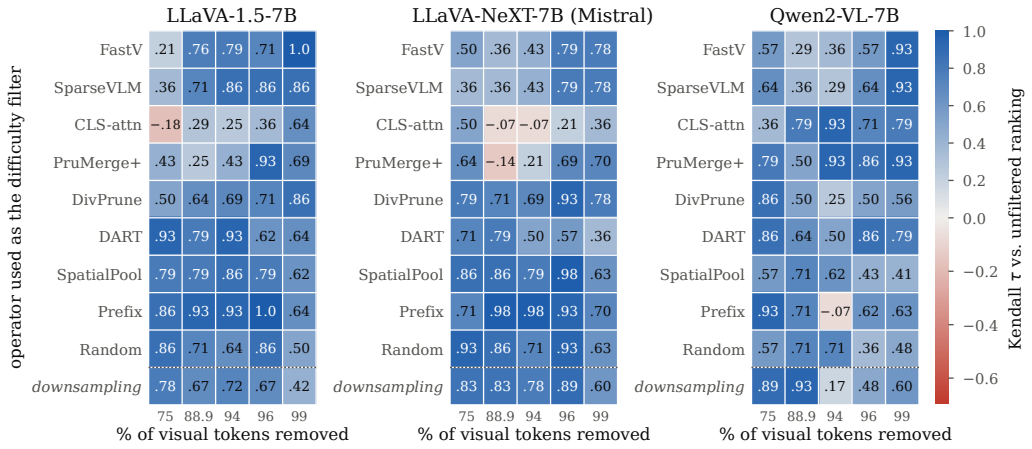

Figure 9: Figure 7 with the filter's own row left out of both rankings, which removes the mechanical part of the distortion (the filter's 0 against competitors that are not at 0). The means are $+0.67$, $+0.62$ and $+0.62$ against a one-dimensional prediction of $+0.85$.