CORRECT AFTER REARRANGEMENT: A TILE-PERMUTATION TEST OF SPATIAL RELIANCE IN VISION-LANGUAGE BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models are scored on benchmarks such as MMStar, RealWorldQA, and ChartQA, where a score above a text-only baseline is often read as evidence that the model used the spatial arrangement of the image. Removing the image shows which scores need an image, not which properties of that image they need. A model could use the spatial arrangement of the scene, or collect evidence from separate regions independently of their layout. We separate these explanations by cutting each image into equal tiles, permuting the tiles at random, and passing the result to the model's own processor, which preserves the pixel content of every tile and disrupts inter-tile adjacency and tile location. We introduce persistence, the probability that a correct answer survives this intervention, and measure it for three open vision-language models on these three benchmarks. At a two-by-two grid, correctness is retained on about 65% to 80% of permutation trials among originally correct items, and on 44% to 79% at a four-by-four grid. This persistence exceeds content-removed baselines by 17 to 55 percentage points on a matched scale in all 27 model-benchmark-grid combinations, including every combination at the finest grid. Scrambling pixels inside the tiles while keeping the layout fixed reduces accuracy to near the text-only baseline, which places within-tile structure ahead of tile arrangement in this ordering. Most originally correct predictions therefore remain recoverable after the original coarse arrangement is disrupted. We release per-item records so that spatial reliance can be examined item by item.

1 INTRODUCTION

Benchmarks such as MMStar (Chen et al., 2024b), RealWorldQA (xAI, 2024), and ChartQA (Masry et al., 2022) are widely used to evaluate the visual capabilities of vision-language models (VLMs). When a model scores well above a text-only baseline, the result is often described as evidence that the model understands the image, and in particular how its contents are arranged. Spatial understanding of this kind requires sensitivity to relations among visual elements beyond their presence (Liu et al., 2023a; Kamath et al., 2023). Benchmark scores alone, however, do not show which aspect of image structure supports the prediction.

The unresolved question is how much of the gain over the text-only baseline depends on spatial arrangement. Text-only comparisons are widely used (Goyal et al., 2017; Agrawal et al., 2018), and some benchmarks, MMStar among them, were built by removing items that a text-only model can answer (Chen et al., 2024b). Such comparisons quantify how much of a score depends on the image at all, and leave open how much of the remaining, image-conditioned score depends on its arrangement. Without this intermediate measurement, two explanations of an above-baseline score remain conflated. The model may be using how things are positioned relative to each other, or gathering local evidence from separate regions without using their arrangement. Both explanations produce the same benchmark score, and only the first requires sensitivity to spatial relations. No existing evaluation separates them on the benchmarks whose scores are cited, because the available controls remove the image instead of rearranging it.

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.

We measure this intermediate condition directly. We partition each benchmark image into a grid of $k \times k$ equal tiles, apply a random permutation to the tile positions, and pass the resulting image through each model's released processor. For items that a model answers correctly on the original image, we measure how often the same model answers correctly after this permutation, and we call this quantity persistence. The permutation conserves visual content, since every pixel appears exactly once both before and after it is applied, so the two inputs differ in where content is placed and not in how much of it is present. Geometry inside a tile is held fixed and only the relation between tiles changes, so persistence measures dependence on arrangement at the tested scale. The grid is a free parameter, so the same measurement runs at the scale a given claim concerns, from the placement of quadrants to the ordering of small local regions.

Persistence is a one-sided diagnostic. When correctness survives, the answer remains recoverable after the original coarse spatial arrangement has been disrupted. A failure after permutation is less diagnostic, because the permutation also introduces boundary discontinuities and object fragmentation that are unrelated to arrangement. We therefore analyze persistence, measured against content-removed baselines on the same items and against a chance-level comparison on the items the model originally missed (Section 5).

Under a uniformly random tile permutation, the expected accuracy of any model is bounded by the best accuracy obtainable from the unordered collection of tiles and the question alone. A coarse tile keeps the arrangement of everything inside it and determines the finer tiles it contains, so this bound cannot increase as the grid becomes finer (Section 3).

We apply this protocol to three publicly released vision-language models from distinct families (LFM2.5-VL-1.6B, SmolVLM2-2.2B-Instruct, and Qwen2.5-VL-3B-Instruct) on the full official MMStar and RealWorldQA evaluation splits and the human-authored ChartQA test subset, using each benchmark's scoring convention and each model's released chat template and processor. Correctness is retained on a majority of permutation trials at the coarsest grid for every model and benchmark, and the gap over content-removed baselines remains large when both quantities are measured on the same items. Control experiments find no significant separator effect after correction, place persistence well above content-removed baselines on a matched scale, and show a sharp accuracy drop when within-tile visual structure is destroyed while the tile grid is left untouched. Persistence declines monotonically with grid size in every model-benchmark pair, the direction the nested-grid ordering predicts, and the rate of decline differs across benchmarks and across MMStar's capability categories, with perception-oriented categories persisting more than reasoning-oriented ones.

We derive an orbit-based upper bound on accuracy under tile permutation and show that the information it bounds is nested across grid sizes, which turns grid granularity into a scale on which spatial reliance can be measured. The construction yields a model-agnostic pixel-space intervention that separates local visual evidence from coarse spatial arrangement within the same image, rather than removing the image, and requires no access to internal representations. We then evaluate three VLMs on three established benchmarks, with matched content-removed baselines, separator controls, and a resolution control, and we release per-item transition records for item-level analysis of spatial reliance. Recovery on originally incorrect items falls below chance in every case, so the retained correctness is not a by-product of guessing. The test requires no new annotation and no access to the model beyond running it again on perturbed images, so its cost is negligible next to running the benchmark itself.

2 Related work

Does the image matter? A long line of work asks whether the visual input is needed at all. VQAv2 (Goyal et al., 2017) pairs similar questions with complementary images, VQA-CP (Agrawal et al., 2018) shifts answer priors to expose question-only shortcuts, and MMStar (Chen et al., 2024b) applies the same idea to modern multimodal benchmarks by identifying samples a text-only model can already answer. The Perceptual Score (Gat et al., 2021) measures how much a multimodal predictor relies on each input, and VisNec (Dong et al., 2026) scores visual necessity during instruction tuning. Closest to our setting, a recent credit-assignment study (Liu et al., 2026b) separates correctness from evidence that the image contributed, using text-only controls and replacement with another image from the same split. These comparisons set an image against no image or against a different image. We keep the same image and test which of its properties the correct answer depends on. To

our knowledge this is the first measurement of how much of an official multimodal benchmark score depends on the arrangement of an image rather than on its content.

Spatial reasoning in vision-language models. Targeted benchmarks show that strong aggregate scores do not imply reliable spatial understanding. VSR (Liu et al., 2023a) and What's Up (Kamath et al., 2023) isolate relations such as *left of* and *above* and report large gaps to human performance, TopViewRS (Li et al., 2024) separates recognition from relative position, and BLINK (Fu et al., 2024), Jigsaw-Puzzles (Lyu et al., 2025), and OmniSpatial (Jia et al., 2026) extend this to perception and higher-order spatial tasks; a recent survey collects the area (Liu et al., 2026a). A parallel line shows that models can behave like bags of words (Yüksekgönül et al., 2023; Thrush et al., 2022) and that some benchmarks permit textual shortcuts (Hsieh et al., 2023), and others improve spatial competence through supervision or architecture (Chen et al., 2024a; Cheng et al., 2024; Ranasinghe et al., 2024; Tong et al., 2024). These works build new tasks or new models, while we intervene on benchmarks already in use and ask what their existing scores rest on.

Patch-level evidence and permuted inputs. In recognition, BagNet (Brendel & Bethge, 2019) reaches high ImageNet accuracy by aggregating evidence from small local patches with little global integration, and texture bias (Geirhos et al., 2019) shows that accuracy can come from cues other than the ones observers assume. For vision transformers (Dosovitskiy et al., 2021), accuracy survives occlusion and spatial perturbation (Naseer et al., 2021), tolerates many patch corruptions (Gu et al., 2022), and holds when patches are dropped during training (Liu et al., 2023b). Qi et al. (2025) bring this to multimodal models by permuting the order of internal visual tokens and report small changes in benchmark scores, and Directed-Tokens (Truong et al., 2025) adds order reconstruction as a training objective. Our intervention differs in three ways. It acts on decoded pixels before each model's released processor, so the same transformation applies across models and no internal representation is modified. It conditions on items the model already answers correctly, yielding a paired within-item comparison across conditions. It also varies grid granularity, which the nested-grid argument of Section 3 ties to the information left after permutation, so the measurement has a scale instead of a single perturbation strength. That argument draws on permutation-invariant set functions (Zaheer et al., 2017; Cohen & Welling, 2016) and on relational reasoning (Santoro et al., 2017), while the evaluated predictors themselves need not be invariant or equivariant.

3 Background and problem setting

3.1 Notation and setup

Consider a benchmark item consisting of a decoded RGB image $I \in [0, 255]^{H \times W \times 3}$, an accompanying question Q, and a ground-truth answer Y. We reflection-pad I on its bottom and right edges so that both dimensions become divisible by 8, producing a padded image $X \in [0, 255]^{H' \times W' \times 3}$. Reflection padding mirrors nearby pixels at the border, so it does not introduce a uniform band, and it is applied identically in every condition.

For a grid parameter $k \in \{2, 4, 8\}$, we partition X into k^2 equal tiles arranged in a $k \times k$ grid, each of size $(H'/k) \times (W'/k)$, which the padding makes exact for every k we use. Let $G_k \cong S_{k^2}$ be the symmetric group acting on the k^2 tile positions. For a permutation $g \in G_k$, $g \cdot X$ denotes the image obtained by placing the tile originally at position i into position $g(i)$, for every i. The three grids sweep the granularity of the relations that are disrupted, from the placement of quadrants at $k{=}2$ to the ordering of small local regions at $k{=}8$, and each coarse tile is exactly a union of finer tiles.

Definition 1 (Orbit statistic)**.** *For a tiled image X, the orbit statistic $O_k(X)$ is the multiset of the k^2 exact pixel tiles of X, with positional labels discarded.*

A predictor h maps a processed image and a question Q to an answer, written $h(P(\cdot), Q)$, where $P(\cdot)$ is a model's released image pre-processing pipeline (resizing, normalization, patchification (Dosovitskiy et al., 2021)). We write $\ell(h(P(\cdot), Q), Y) \in \{0, 1\}$ for the benchmark's correctness score after answer extraction. We let D denote the empirical distribution of items (X, Q, Y) induced by a benchmark split; all expectations below are taken jointly over $(X, Q, Y) \sim D$ and independent randomness in the group action.

3.2 INFORMATION BOUNDS UNDER TILE PERMUTATION

The group-averaging argument uses a predictor that sees only the orbit statistic $O_k(X)$ and the question Q, and not the positionally labeled image X, in the sense used for functions of unordered collections (Zaheer et al., 2017; Cohen & Welling, 2016).

Definition 2 (Symmetrized predictor)**.** *Given the orbit statistic $o = O_k(X)$ and the question q, let $\bar{h}(o, q)$ denote the randomized predictor that samples $g' \sim \mathrm{Uniform}(G_k)$ independently of everything else, reconstructs an image X' by placing the tiles of o according to g', and outputs $h(P(X'), q)$. The distribution of $\bar{h}(o, q)$'s output depends only on (o, q), not on which particular arrangement of X gave rise to o, so $\bar{h}$ is measurable with respect to $(O_k(X), Q)$.*

Definition 3 (Orbit Bayes accuracy)**.** *For a grid k and distribution D over items, $\mathrm{BayesAcc}_D(O_k, Q)$ is the supremum, over all (possibly randomized) predictors f that are measurable with respect to $(O_k(X), Q)$, of $\mathbb{E}_{(X,Q,Y)\sim D}\big[\ell(f(O_k(X), Q), Y)\big]$.*

Proposition 1 (Group-averaging bound)**.** *Let $(X, Q, Y) \sim D$ and let g be drawn uniformly from G_k, independently of (X, Q, Y). Then*

$$\mathbb{E}_{(X,Q,Y),g}\big[\ell(h(P(g\cdot X), Q), Y)\big] = \mathbb{E}_{(X,Q,Y)}\big[\mathbb{E}_{\bar{h}}[\ell(\bar{h}(O_k(X), Q), Y)]\big] \;\le\; \mathrm{BayesAcc}_D(O_k, Q).$$

The equality holds because averaging $h(P(g\cdot X), Q)$ over a uniformly random g, for a fixed (X, Q), is by construction exactly the distribution of $\bar{h}(O_k(X), Q)$'s output. The inequality holds because $\bar{h}$ is one particular predictor measurable with respect to $(O_k(X), Q)$, and $\mathrm{BayesAcc}_D(O_k, Q)$ is the supremum accuracy over that class of predictors. In words, the expected accuracy of any model under a uniformly random tile permutation cannot exceed the best possible accuracy obtainable from the unordered tile content and the question alone, with no information about the original adjacency or absolute tile locations, computed over the benchmark's item distribution.

The bound applies to the class of predictors that see the orbit statistic and the question, and the appendix examines how tight it is on a fixed evaluation split. The next proposition orders that class across grid sizes.

Proposition 2 (Direction of the nested-grid bound)**.** *Let $k_1 < k_2$ be nested grids, meaning every k_1-tile is exactly a union of k_2-tiles. Because each k_1-tile carries its full pixel content, it can be deterministically subdivided into its constituent k_2-tiles using a fixed rule that does not depend on the k_1-tile's position. Applying this subdivision to every tile in the multiset $O_{k_1}(X)$ recovers the multiset $O_{k_2}(X)$, so there exists a fixed deterministic map φ with $O_{k_2}(X) = \varphi(O_{k_1}(X))$ for every X. Consequently, $O_{k_2}(X)$ is measurable with respect to $O_{k_1}(X)$, so any predictor measurable with respect to (O_{k_2}, Q) is also measurable with respect to (O_{k_1}, Q), and*

$$\mathrm{BayesAcc}_D(O_{k_2}, Q) \;\le\; \mathrm{BayesAcc}_D(O_{k_1}, Q).$$

The orbit-based bound therefore cannot increase as the grid becomes finer.

The orbit statistics are therefore nested across grid sizes. The two propositions together name the information the intervention removes and order it across grids, which is what lets the grid sweep be read as a monotone reduction of positional information. The bound concerns raw accuracy under uniform permutations, while the behavioral quantity we report is persistence under derangements, so we report them side by side and not as estimates of each other.

3.3 METRICS

For a condition c (for example, the original image, the identity condition, or a permuted image at grid k), raw accuracy is

$$\mathrm{Acc}_c = \frac{1}{N}\sum_{i=1}^{N} \ell\big(h(P(c(X_i)), Q_i), Y_i\big).$$

Write $\ell_{\mathrm{orig},i}$ for item i's correctness on the padded original image. Persistence among originally correct items restricts the average to the subset with $\ell_{\mathrm{orig},i} = 1$:

$$\mathrm{Pers}_c = \frac{1}{|\{i : \ell_{\mathrm{orig},i} = 1\}|} \sum_{i:\ell_{\mathrm{orig},i}=1} \ell\big(h(P(c(X_i)), Q_i), Y_i\big),$$

and gain among originally incorrect items is defined analogously on the complementary subset. Persistence is our primary quantity because it restricts attention to items the model already answers correctly on the original image. Each measurement then compares the same item across conditions, and each model is evaluated relative to its own originally correct set, which places models and benchmarks with very different accuracies on a single scale. Because persistence conditions on a subset of items, it is not on the same scale as an unconditioned, population-level accuracy such as raw blind accuracy. Our scale-matched comparator is the target-specific persistence gap, which subtracts, for each originally correct item, the mean correctness of the content-removed baselines from the correctness of the permuted image of that same item,

$$\mathrm{TSP}_c = \frac{1}{|\{i : \ell_{\mathrm{orig},i} = 1\}|} \sum_{i:\ell_{\mathrm{orig},i}=1} \Big(\ell\big(h(P(c(X_i)), Q_i), Y_i\big) - \bar{\ell}_i^{\mathrm{cf}}\Big),$$

where $\bar{\ell}_i^{\mathrm{cf}}$ is item i's mean correctness over the content-removed baselines.

4 METHOD

4.1 INTERVENTION

The permutation acts on the padded image before it reaches any model-specific processor. The same pixel-space transformation is therefore applied for every model, each model receives the perturbed image through its released preprocessing, and model internals are left unchanged. Every comparison also uses a canvas of identical size in both conditions, so the permutation leaves unchanged the resolution at which a model processes the image. For every item and every grid $k \in \{2, 4, 8\}$ we build three conditions, an identity condition that reassembles the tiles in their original order, a permuted condition, and a separated permuted condition that adds a mean-color separator at each internal tile boundary. The identity condition carries the reassembly without the permutation, which separates the effect of moving tiles from the effect of rebuilding the image. The separator is the only visual element that reassembly introduces, so comparing the separated and separator-free conditions under the same permutation isolates its contribution. The persistence values in the main text use the separator-free condition.

Two kinds of random permutation are sampled per item and grid. Uniform permutations are drawn from the full symmetric group S_{k^2} with fixed points allowed, which is the case Proposition 1 bounds. Derangements are fixed-point-free and are used for the persistence values in the main text, so every tile changes its absolute position under the reported intervention.

A $k{=}2$ derangement moves every tile, and derangements still differ in how much tile adjacency they leave intact. We stratify them by retained adjacency and report that analysis in the appendix.

We also construct a $k{=}2$ within-tile shuffle that permutes pixels inside each tile while leaving the tiles in place, which destroys the spatial structure inside a tile and preserves its set of pixel values. The two interventions perturb complementary aspects of spatial structure, since one keeps the content of every tile and changes where it sits, while the other keeps every tile in place and destroys its internal organization.

4.2 BASELINES

Blind evaluation presents the same question without the image, using each model's text-only interface, so it measures question-only performance. Random-image replacement substitutes a dimension-matched image from the same benchmark split (Liu et al., 2026b), breaking the correspondence between image and question while retaining an image input. No item is ever paired with its own image. We refer to these two conditions as content-removed baselines, since both remove the content of the item's own image.

4.3 MODELS

We evaluate three vision-language models without fine-tuning or architectural modification: LFM2.5-VL-1.6B (Amini et al., 2025), SmolVLM2-2.2B-Instruct (Marafioti et al., 2025), and Qwen2.5-VL-3B-Instruct (Bai et al., 2025), abbreviated below and in the tables as LFM, Smol,

and Qwen. Each model uses its released preprocessing and inference pipeline, with the full configuration given in the appendix. Generation uses deterministic greedy decoding, and token budgets are held fixed across conditions. Qwen uses a dynamic-resolution processor that allocates more visual tokens to a larger input, while LFM and Smol allocate a fixed number regardless of input size. The resolution control of Section 5 therefore applies to Qwen.

4.4 Datasets and scoring

We evaluate the benchmarks whose scores are cited in support of claims about spatial understanding, choosing three that expect arrangement to matter for different reasons. RealWorldQA (xAI, 2024) is built around real-world spatial judgments, MMStar (Chen et al., 2024b) keeps only items that require the image, and ChartQA (Masry et al., 2022) encodes meaning in the arrangement of marks, axes, and labels. We use the complete official MMStar validation split (1,500 items with six capability categories), the complete official RealWorldQA evaluation split (765 items), and the human-authored subset of the ChartQA test split (1,250 items). We implement the published scoring convention of each benchmark, including ChartQA's relaxed numeric comparator with its 5% relative tolerance (Masry et al., 2022). The extraction rules, dataset revisions, and failure diagnostics are given in the appendix.

4.5 Uncertainty and multiplicity

Each statistic is paired with a bootstrap scheme that matches its structure. Quantities averaged over several permutations per item use a two-level percentile bootstrap that resamples items and then resamples that item's permutations, and quantities with a single observation per item use an item-level bootstrap. Each scheme is stated in full in the appendix.

MMStar bootstraps resample items within each capability category, since category membership is a designed stratification of that benchmark. Holm correction is applied to the two families of hypothesis tests, the 27 separator-line contrasts and the 9 retention-stratified contrasts, and intervals for descriptive summaries are reported without multiplicity adjustment.

5 Results

5.1 Persistence and matched-scale baselines

Table 1 reports native and padded accuracy, separator-free persistence at $k{=}2, 4, 8$, and the blind, random-image, and within-tile-shuffle control accuracies for every model and benchmark. Padding changes native accuracy by less than one point in every model-benchmark pair except Qwen on RealWorldQA, where it is 2.22 points higher, so the padded condition reproduces the benchmark score that the intervention is compared against.

At $k{=}2$, persistence ranges from 64.99% (LFM on ChartQA) to 80.36% (Smol on RealWorldQA) across the nine model-benchmark pairs. The blind and random-image baselines in Table 1 are computed over the full item set. We therefore restrict them to the same originally correct subset (Table 3) and compute the target-specific persistence gap defined in Section 3, which compares each item against its own content-removed reference. Qwen's $k{=}2$ persistence gap on ChartQA is $+54.81$ percentage points (pp, 95% confidence interval $[+52.24, +57.40]$), and every one of the 27 model-benchmark-grid combinations is positive with a confidence interval excluding zero, ranging from about $+17$ to $+55$ percentage points (Table 6). The gap holds at the finest grid as well, where it stays at or above $+17$ pp in every combination and persistence remains above half of trials on MMStar and RealWorldQA. Within each benchmark the three models fall in a narrow band of persistence values, so the level of persistence follows the benchmark more closely than the model, which is what a property of the evaluation items rather than of one architecture would produce. This comparison is available because the intervention is applied in pixel space, so models with different visual tokenizers receive the same transformation and can be read on one axis.

Across every model and benchmark, permuted accuracy lies between the padded and blind values and closer to padded, and the average over uniform permutations tracks the average over derangements closely. Figure 2 in the appendix compares all six conditions directly.

Table 1: Main results. Native accuracy is the benchmark score and padded accuracy is the intervention reference. Persistence columns are separator-free persistence on originally correct (padded) items. The control columns give blind, random-image, and within-tile-shuffle accuracy as raw, population-level quantities over all items, on a different scale from the persistence columns. All values in percent.

		Original		Persistence			Controls		
Model	Benchmark	Native	Padded	k=2	k=4	k=8	Blind	Random	Within-tile
LFM	MMStar	47.00	47.20	71.24	62.15	53.67	27.27	25.00	27.09
LFM	RealWorldQA	63.79	64.18	76.33	71.83	65.92	36.86	36.91	34.99
LFM	ChartQA	66.80	66.40	64.99	43.94	33.21	12.80	12.45	13.12
Smol	MMStar	43.20	43.93	71.59	64.14	57.97	26.00	24.80	25.53
Smol	RealWorldQA	57.39	57.39	80.36	78.82	74.26	33.33	39.00	39.48
Smol	ChartQA	51.28	50.80	66.20	51.97	40.26	10.48	14.72	14.16
Qwen	MMStar	54.60	54.40	70.64	62.01	53.68	28.33	27.44	30.76
Qwen	RealWorldQA	61.18	63.40	73.48	69.90	63.37	39.74	35.64	34.25
Qwen	ChartQA	71.28	70.96	70.71	51.71	36.08	12.88	13.36	14.45

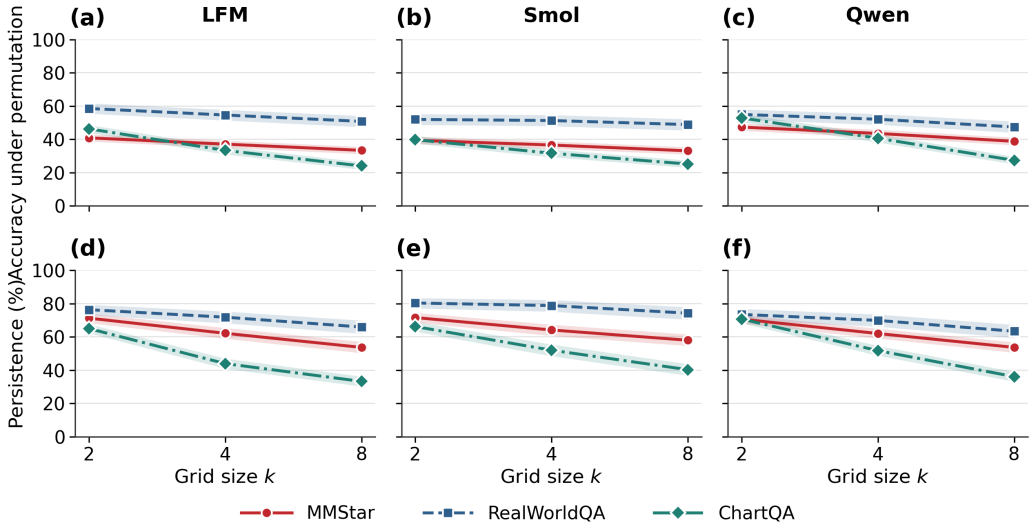


Figure 1: Raw accuracy under permutation and persistence against grid size k, one column per model and one curve per benchmark. (a) to (c) accuracy under 5 (k=2) or 3 (k=4, 8) separator-free permutations drawn from the full permutation group, with fixed points allowed. (d) to (f) persistence under 5 (k=2) or 3 (k=4, 8) separator-free derangements. Shaded bands are two-level bootstrap 95% intervals, category-stratified for MMStar.

5.2 Persistence across grid sizes

Figure 1 plots two quantities against $k \in \{2, 4, 8\}$, one column per model and one curve per benchmark. The top row shows raw accuracy under uniform permutations, which is the quantity bounded by Proposition 1, and the bottom row shows separator-free persistence under derangements. For all nine model-benchmark pairs, persistence decreases monotonically from k=2 to k=8, and it stays above half of trials at k=4 in eight of the nine pairs. The rate of decline differs across benchmarks, with RealWorldQA declining the least (from about 73 to 80% at k=2 down to 63 to 74% at k=8), MMStar in between (from about 70 to 72% down to 54 to 58%), and ChartQA the fastest (from about 65 to 71% down to 33 to 40%). The steeper decline on ChartQA is consistent with its dependence on the arrangement of marks, axes, and labels (Masry et al., 2022), a property it shares with chart and plot tasks more generally (Kafle et al., 2018; Methani et al., 2020).

5.3 LOCAL CONTENT VERSUS TILE ARRANGEMENT

The within-tile shuffle, which permutes pixels inside each $k{=}2$ tile while leaving the tiles in place, reduces raw accuracy to between 13.12% and 39.48% across the nine model-benchmark pairs (Table 1, rightmost column), near the blind and random-image baselines. Uniform permutations, measured on the same population-level scale (Table 2), retain substantially more accuracy. On MMStar, for example, accuracy under uniform $k{=}2$ permutation ranges from 39.47% to 47.39% across the three models, well above both the within-tile-shuffle accuracy (25.53% to 30.76%) and blind accuracy (26.00% to 28.33%) on the same benchmark. Within-tile pixel scrambling therefore produces a substantially larger reduction in accuracy than whole-tile permutation. The two interventions remove opposite parts of the input, and the observed drop is far larger when the content inside the tiles is destroyed than when intact tiles are moved, which places within-tile structure ahead of tile arrangement in the ordering these measurements support, in line with recognition from local patch evidence alone (Brendel & Bethge, 2019).

The separator-line ablation compares separated and separator-free permutations that reuse identical permutation orders, so it isolates the effect of inserting a one-pixel mean-color line at internal tile boundaries. None of the 27 contrasts remains significant after Holm correction; the smallest Holm-adjusted p-value is 0.065, and the point estimates span -1.79 to -0.76 pp (Table 7 in the appendix).

5.4 PERSISTENCE BY MMSTAR CAPABILITY CATEGORY

We break MMStar down into its six official capability categories (Chen et al., 2024b) and compare original accuracy with separator-free $k{=}2$ persistence (Figure 3 in the appendix). Persistence varies across categories, with coarse perception and math persisting highest across all three models (76% to 79%). Fine-grained perception is the lowest or near-lowest category for every model (59% to 61%). This pattern is compatible with coarse visual attributes being recoverable from local evidence, while fine-grained comparisons that relate small details across the image are more sensitive to the intervention.

5.5 COMPARISON WITH A LABEL-RANDOMIZATION NULL

On MMStar we compare the observed gain against an item-level label-randomization null under which the gold option is uniform over each item's verified option count, so each trial is Bernoulli$(1/K_i)$. Observed gains sit below this null in every one of the nine model-grid combinations, with a two-sided Monte Carlo p-value of 0.0002 in all cases, below a Bonferroni-corrected threshold (Figure 4 in the appendix). Recovery on these items is therefore inconsistent with uniform guessing.

5.6 RESOLUTION CONTROL AT THE FINEST GRID

We next test whether the sharp $k{=}8$ decline can be attributed to reduced effective tile resolution, since a model's image resizing may render small tiles illegible independently of any loss of arrangement. The test applies to Qwen, whose processor allocates more visual tokens to a larger input, while LFM and Smol allocate a fixed number. We upsample each of the 64 $k{=}8$ tiles by a factor of four in each linear dimension before reassembly, and compare identity and permuted accuracy at standard and high resolution.

The resulting difference-in-differences, which measures whether upsampling closes more of the permuted-versus-identity gap than it changes identity accuracy alone, is not distinguishable from zero on MMStar and ChartQA, and is positive on RealWorldQA ($+4.71$ percentage points, 95% confidence interval $[+1.09, +8.41]$). In no benchmark does the added resolution bring the permuted condition back to the identity level, and the ChartQA interval is $[-0.05, +5.31]$, so reduced tile resolution alone is unlikely to account for the $k{=}8$ decline (Figure 6 in the appendix).

6 DISCUSSION

The gap between $k{=}2$ persistence and content-removed baselines on a matched scale is large and replicates across three models from distinct families and three benchmarks. Correct answers are re-

produced on a majority of permutation trials after the coarse spatial arrangement is disrupted. At the item level, 40% to 65% of originally correct items are correct under every sampled $k{=}2$ derangement, and at most 13% under none of them, so persistence is a property of most originally correct items and not of a small permutation-insensitive subset. Sensitivity to the intervention is graded, varying across MMStar capability categories and growing as the grid becomes finer, most visibly on ChartQA, so these benchmarks contain an identifiable subset of items on which the intervention does change the answer. Persistence is highest for coarse perception and math and lowest for fine-grained comparisons that relate small details across the image, an ordering consistent with greater sensitivity where the answer depends on comparing separate regions. The retention-stratified analysis in the appendix provides convergent evidence, with persistence higher in the high-retention stratum than in the zero-retention stratum in all nine pairs, by $+3.21$ to $+16.91$ percentage points and with no pair showing the reverse ordering. Adjacency that survives a permutation therefore matters in a graded way, which is the same ordering the grid sweep produces at a coarser resolution.

Two controls address artifacts introduced by the intervention. The separator-line ablation indicates that the added mean-color line at tile boundaries is unlikely to be a major driver of the measured persistence, and the resolution control indicates that reduced tile resolution is unlikely to account for the $k{=}8$ decline on two of the three benchmarks.

Persistence is a behavioral diagnostic and does not by itself reveal the mechanism used on the original image. The analyses in Section 5 and the appendix are inconsistent with uniform guessing on originally incorrect items and with correctness driven by language priors alone. A model may also partially infer the arrangement from the content of individual tiles, although models are reported to be weak at recovering the order of shuffled image regions (Lyu et al., 2025). Distinguishing such reconstruction from answerability that does not need arrangement calls for mechanistic analysis.

A benchmark score above the text-only baseline shows that the image contributes at the aggregate level, and the tile-permutation test supplies the check that the score itself cannot provide, namely whether the arrangement of that image was needed. Because the test reuses the benchmark's own items and scoring rule, it speaks to the score that is cited and not to a separate dataset assembled for the purpose. High persistence shows that correct predictions remain recoverable after the original coarse layout is disrupted. Low persistence, on its own, is not sufficient evidence of spatial reliance, since the permutation also changes the input in ways unrelated to arrangement, and low-persistence items serve as candidates for more targeted analysis of spatial dependence. Because persistence is defined per item, the released records split a benchmark into the items whose answers survive the intervention and those whose answers do not, without new annotation, which identifies the subset on which a claim about arrangement is actually being tested. Benchmark authors can use it the way text-only filtering is already used (Chen et al., 2024b; Dong et al., 2026), reporting it alongside accuracy and examining high-persistence items separately.

Limitations. Our study covers three open models with 1.6B to 3B parameters and three benchmarks. Whether the measured persistence levels extend to larger or differently trained models remains open (Liu et al., 2026a), and benchmarks written specifically for spatial relations (Liu et al., 2023a; Kamath et al., 2023; Jia et al., 2026) are the natural next application of the same protocol. Because within-tile pixel scrambling is a substantially stronger perturbation than tile permutation, the relative size of the two drops should not be read as a quantitative decomposition of local-content and spatial-layout contributions; interventions of matched severity would be needed for that.

7 Conclusion

We measured how often three vision-language models preserve correct answers after their benchmark images are cut into tiles and the tiles are permuted, an intervention that disrupts inter-tile adjacency and tile location while leaving each tile's pixels untouched. Persistence at the coarsest grid is about 65% or higher for every model and benchmark, the matched-scale gap over content-removed baselines is positive in all 27 combinations with no exception at any grid size, and scrambling pixels within a fixed layout reduces accuracy to near those baselines. The same levels appear in all three models, so persistence tracks the benchmark and not the architecture. Most of the correct predictions these benchmarks record therefore remain recoverable after the coarse spatial arrangement is disrupted, and the test is directly applicable to other models and benchmarks.

AI USE STATEMENT

Generative AI tools were used for language editing and for support in writing the experiment code and the figure-generating scripts. The authors verified all experiments, analyses, and reported values, and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The complete experimental protocol, including grid definitions, padding rules, permutation sampling, decoding parameters, and scoring rules, is described in Section 4, with further implementation detail and diagnostics in the appendix. We pin all model and dataset revisions to exact commit hashes and release every item-level output, sampled permutation, and correctness label, so every number reported in this paper can be recomputed from those records.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, Jacob Marks, Edoardo Mosca, Samuel J. Paech, Paul Pak, Rom N. Parnichkun, Alex Quach, Ryan Rogers, Daniela Rus, Nayan Saxena, Bettina Schlager, Tim Seyde, Jimmy T. H. Smith, Aditya Tadimeti, and Neehal Tumma. LFM2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

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

Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a. doi: 10.1109/CVPR52733.2024.01370.

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*, volume 37, pp. 27056–27087, 2024b. doi: 10.52202/079017-0850.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024. doi: 10.52202/079017-4293.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. VisNec: Measuring and leveraging visual necessity for multimodal instruction tuning. In *Computer Vision – ECCV 2026*, pp. 531–548. Springer Nature, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9.

Itai Gat, Idan Schwartz, and Alexander G. Schwing. Perceptual score: What data modalities does your model perceive? In *Advances in Neural Information Processing Systems*, volume 34, pp. 21630–21643, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

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*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Computer Vision – ECCV 2022*, volume 13672 of *Lecture Notes in Computer Science*, pp. 404–421. Springer, 2022. doi: 10.1007/978-3-031-19775-8_24.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1355.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, 2026.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2018. doi: 10.1109/CVPR.2018.00592.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.106.

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Spatial intelligence in vision-language models: A comprehensive survey. *Artificial Intelligence Review*, 2026a. doi: 10.1007/s10462-026-11671-x.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566.

Feixiang Liu, Qiang Qiu, Lanbo Sun, Nan Wei, Huawei Shen, and Xueqi Cheng. Visual credit audit for multimodal spatial reasoning. *arXiv preprint arXiv:2607.27069*, 2026b.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. PatchDropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3942–3951, 2023b. doi: 10.1109/WACV56688.2023.00394.

Zesen Lyu, Dandan Zhang, Wei Ye, Fangdi Li, Zhihang Jiang, and Yao Yang. Jigsaw-puzzles: From seeing to understanding to reasoning in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25992–26003. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1320.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. In *Conference on Language Modeling (COLM)*, 2025.

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 the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1516–1525, 2020. doi: 10.1109/WACV45572.2020.9093523.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23296–23308, 2021.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024. doi: 10.1109/CVPR52733.2024.01233.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4967–4976, 2017.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024. doi: 10.52202/079017-2771.

Thanh-Dat Truong, Huu-Thien Tran, Thai Son Tran, Bhiksha Raj, and Khoa Luu. Directed-tokens: A robust multi-modality alignment approach to large language-vision models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 155815–155839, 2025. doi: 10.52202/085713-4694.

xAI. RealWorldQA, 2024. Benchmark dataset.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

A Appendix

A.1 Tightness of the group-averaging bound

Proposition 1 states an upper bound over predictors measurable with respect to the orbit statistic and the question. On a finite evaluation split that supremum need not be a tight empirical characterization of spatial information, since exact pixel tiles are rarely duplicated across items and the orbit statistic can identify individual items at fine grids. We therefore read the propositions as an ordering of the orbit information across nested grids, which holds for every predictor class.

A.2 Resolution control conditions

Both conditions of the resolution control use separated permutations, because the high-resolution mosaic inserts a separator line at tile boundaries after each tile is upsampled, so the standard-resolution comparison is taken from the matching separated $k{=}8$ condition.

A.3 Bootstrap schemes

Persistence, raw accuracy under permutation, and gain, all of which involve several permutation samples per item, use a two-level percentile bootstrap with $10{,}000$ resamples; items are resampled with replacement, and, for each resampled item, the full set of available permutation samples for that item (5 at $k{=}2$, 3 at $k{=}4$ and $k{=}8$) is itself resampled with replacement and averaged, matching how the point estimate averages over the same samples. Native, padded, and blind accuracy, and paired differences that involve a single observation per item, use a one-level item bootstrap. Random-image statistics, which involve three independent replacement maps per item, use an item-and-map two-level bootstrap. The retention-stratified analysis uses a nested design bootstrap that resamples items and then redraws each item's within-stratum outcomes. For MMStar, the persistence, permuted-accuracy, and simple-accuracy bootstraps resample items within each capability category, preserving its item count; the remaining statistics are computed without this stratification.

A.4 Implementation of the intervention

The separator overwrites, at each internal grid line, the topmost row of pixels of every tile below an internal horizontal line and the leftmost column of pixels of every tile to the right of an internal vertical line, using the channel-wise mean color of the decoded image. At $k{=}2$, the five permutations of an item are drawn jointly and uniformly without replacement from the 24 elements of the full group and from the 9 derangements of S_4, which the size of these spaces allows. At $k{=}4$ and $k{=}8$, three permutations of each kind are drawn per item with deterministic seeds indexed by item and sample. Retained adjacency after a $k{=}2$ derangement defines zero-, low-, and high-retention strata with population shares $3/9$, $4/9$, and $2/9$.

Each model is loaded with bfloat16 weights on a single GPU. Generation is deterministic greedy decoding, and the token budget is 8 tokens for MMStar, 16 for RealWorldQA, and 32 for ChartQA, held constant across conditions within each benchmark.

A.5 Extraction rules

MMStar answers are scored by taking the last isolated letter among A through D in the model's output as the model's choice and comparing it against the gold letter. RealWorldQA answers are scored by normalized text equality after removing articles and punctuation. ChartQA answers are scored by a relaxed numeric comparator with a 5% relative tolerance, applied after common answer prefixes are removed and a single numeric span, with an optional percent suffix, is located. An empty output is scored as an extraction failure for every benchmark. A non-empty output containing no numeric span is scored as an ordinary textual answer by normalized string equality against the gold value. Only non-empty outputs whose cleaned text contains more than one numeric span, or a single span that cannot be parsed under the percent-suffix convention, are counted as extraction failures.

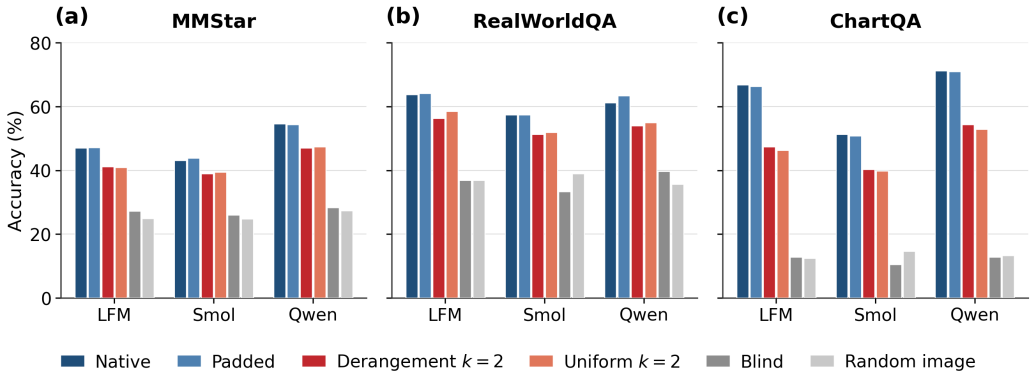


Figure 2: Accuracy per condition for each model and benchmark, with (a) MMStar, (b) RealWorldQA, and (c) ChartQA. Conditions are the native original, the padded original (intervention reference), separator-free derangements at k=2 averaged over 5 samples, uniform full-group k=2 averaged over 5 permutations, blind text-only accuracy, and the mean over 3 independent dimension-matched random-image maps.

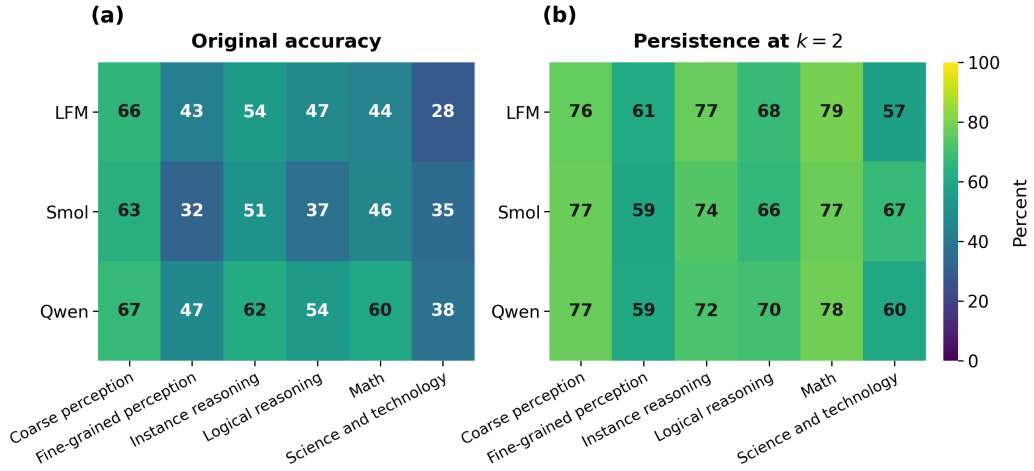


Figure 3: MMStar capability-category breakdown, with about 250 items per category. (a) padded-original accuracy per category and model. (b) separator-free persistence at k=2 per category. Both panels report percentages on the shared color scale.

A.6 Condition decomposition

Figure 2 shows all six conditions side by side for each model and benchmark, from the native and padded originals through separator-free derangements at k=2 averaged over five samples, the uniform k=2 average, blind accuracy, and the mean over three random-image maps.

A.7 Capability-category breakdown

Figure 3 reports MMStar accuracy and persistence for each of the six official capability categories.

A.8 Gain relative to the label-randomization null

Figure 4 compares the observed gain on originally incorrect MMStar items against the item-level label-randomization null described in Section 5, for every model and grid size.

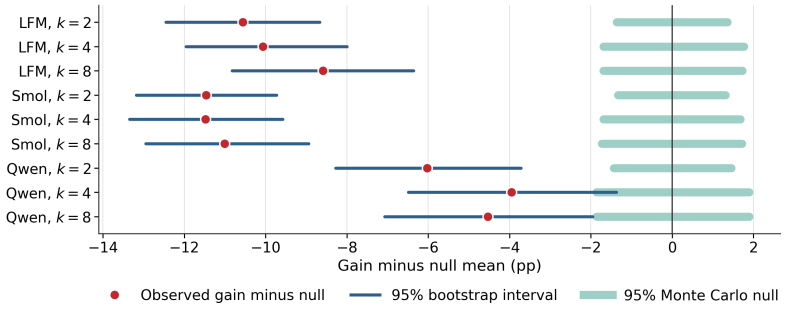


Figure 4: MMStar gain on originally incorrect items relative to an item-level label-randomization null (gold option uniform over each item's verified option set), one row per model and grid size. Points are the gain minus the null mean, dark segments are item-cluster 95% bootstrap intervals of this difference, and light bands are Monte Carlo 95% intervals of the null itself. The two-sided Monte Carlo p-value is 0.0002 for every row.

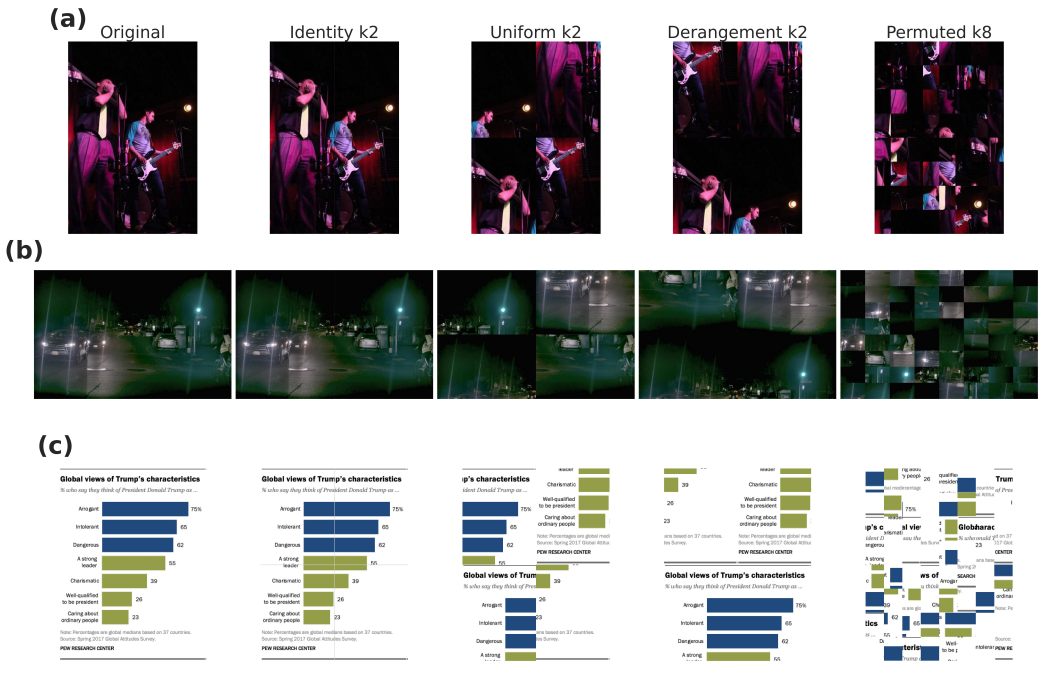


Figure 5: Example inputs (item index 7 of each benchmark; (a) MMStar, (b) RealWorldQA, (c) ChartQA), showing the pixel inputs supplied to the released processors. From left to right in each row are the padded original, the identity reassembly at k=2 with the one-pixel internal separator, a sampled uniform full-group k=2 permutation, a sampled derangement at k=2, and a permuted k=8 mosaic.

A.9 Example inputs

Figure 5 shows the pixel inputs delivered to each model's processor for one illustrative item from each benchmark, from the padded original through the identity reassembly with its one-pixel separator, a sampled uniform permutation, a sampled derangement, and a permuted k=8 mosaic. Each tile at k=2 remains intact and recognizable, whereas k=8 fragments objects and text into small, disconnected pieces.

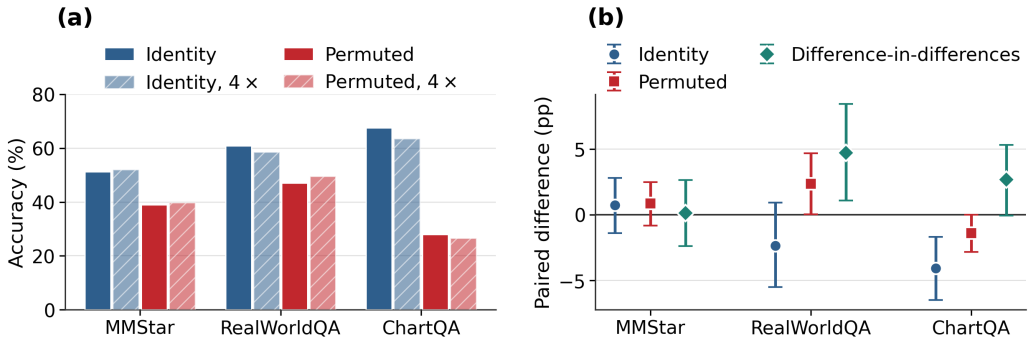


Figure 6: Qwen-only resolution control at k=8, using separated permutations in both conditions. (a) identity and permuted accuracies at standard and $4\times$ resolution. (b) item-paired high-resolution minus standard-resolution contrasts with 95% bootstrap intervals, showing the identity change, the permuted change, and the difference-in-differences. LFM and Smol are excluded because their processors do not receive a larger pixel grid under the $4\times$ mosaic.

Table 2: Raw accuracy under uniform full-group separator-free permutations (percent), together with padded-original accuracy.

Model	Benchmark	Padded acc	Uniform k=2	Uniform k=4	Uniform k=8
LFM	MMStar	47.20	40.92	37.13	33.40
LFM	RealWorldQA	64.18	58.56	54.64	50.81
LFM	ChartQA	66.40	46.32	33.36	24.03
Smol	MMStar	43.93	39.47	36.60	33.11
Smol	RealWorldQA	57.39	52.00	51.33	48.89
Smol	ChartQA	50.80	39.82	31.65	25.17
Qwen	MMStar	54.40	47.39	43.56	38.82
Qwen	RealWorldQA	63.40	54.98	52.11	47.45
Qwen	ChartQA	70.96	52.88	40.67	27.36

A.10 Resolution control

Figure 6 reports the Qwen-only resolution control at k=8 described in Section 5.

A.11 Raw accuracy under uniform permutations

Table 2 reports raw accuracy under permutations drawn from the full symmetric group S_{k^2}, including permutations with fixed points, which is the quantity bounded by Proposition 1; the corresponding trend across k is shown in Figure 1(a) to (c) in the main text. The two quantities track each other closely across all nine model-benchmark pairs.

A.12 Matched-scale content-removed baselines

Table 3 reports identity-conditioned persistence, which measures the same permuted condition while conditioning on the items answered correctly under the identity condition instead of the original image, together with the blind and random-image accuracies restricted to the originally correct subset. Identity-conditioned persistence stays within a few points of the values in Table 1, so the measurement does not depend on which of the two unpermuted conditions defines the conditioning set. The conditioned baselines are the scale-matched comparator for persistence and the basis of Table 6.

Table 3: Identity-conditioned persistence, that is, separated permuted persistence among items answered correctly under the identity condition, together with the blind and random-image baselines restricted to originally correct items. All columns are conditional accuracies in percent.

Model	Benchmark	Identity-conditioned	Blind	Random image
LFM	MMStar	70.97	38.98	33.43
LFM	RealWorldQA	76.69	41.14	44.13
LFM	ChartQA	64.58	16.02	16.10
Smol	MMStar	73.10	37.18	37.43
Smol	RealWorldQA	78.24	42.14	57.25
Smol	ChartQA	66.55	16.54	22.99
Qwen	MMStar	71.52	35.78	34.80
Qwen	RealWorldQA	75.90	48.04	43.99
Qwen	ChartQA	70.76	15.90	15.90

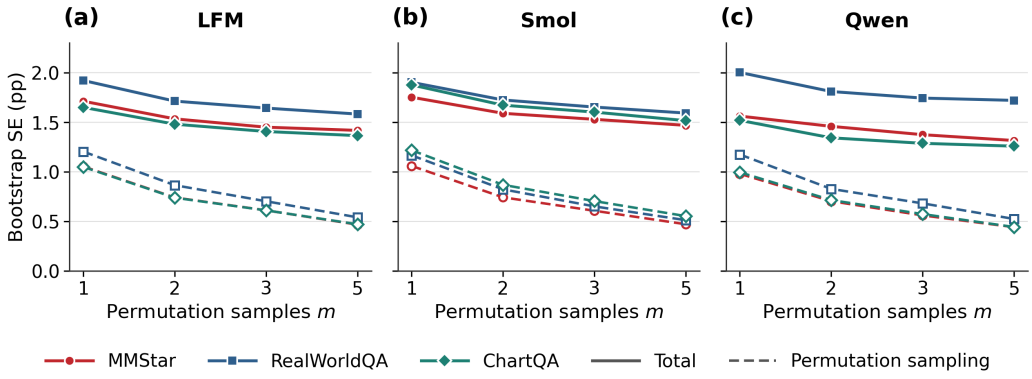


Figure 7: Uncertainty decomposition of the separator-free k=2 persistence estimate against the number m of permutation samples averaged per item, with (a) LFM, (b) Smol, and (c) Qwen. Solid lines with filled markers show the total two-level bootstrap standard error and dashed lines with open markers show the permutation-sampling component with the item set fixed.

A.13 Sensitivity to the number of permutation samples

The persistence estimate averages five permutation samples per item at k=2 and three at k=4 and k=8. Figure 7 and Table 4 decompose the total bootstrap standard error into an item-resampling component and a permutation-sampling component, as a function of the number m of samples averaged per item at k=2. The permutation-sampling component is smaller than the total for every model and benchmark and falls by more than half between $m = 1$ and $m = 5$, so the reported intervals are dominated by item-level variation.

A.14 Conditional transitions and per-item spectrum

Figure 8 reports two views of the separator-free k=2 condition. The first gives the share of originally correct trials that remain correct and the share of originally incorrect trials that become correct, each normalized within its own outcome stratum. The second gives the distribution of individual items into those that are never, sometimes, or always correct across the five sampled derangements. Across all nine model-benchmark pairs, between 40% and 65% of originally correct items are correct under every one of the five sampled permutations, and at most 13% are never correct under any of them.

Table 4: Bootstrap standard error of separator-free k=2 persistence as a function of the number m of permutation samples averaged per item, for every model and benchmark. Total SE is the two-level bootstrap standard error and permutation SE is the component from permutation sampling with the item set fixed. All standard errors in pp.

		$m = 1$		$m = 5$	
Model	Benchmark	Total SE	Perm. SE	Total SE	Perm. SE
LFM	MMStar	1.71	1.05	1.42	0.47
LFM	RealWorldQA	1.92	1.20	1.58	0.54
LFM	ChartQA	1.65	1.05	1.36	0.47
Smol	MMStar	1.75	1.06	1.47	0.47
Smol	RealWorldQA	1.90	1.17	1.59	0.51
Smol	ChartQA	1.88	1.22	1.52	0.55
Qwen	MMStar	1.56	0.98	1.32	0.44
Qwen	RealWorldQA	2.00	1.17	1.72	0.52
Qwen	ChartQA	1.52	1.00	1.26	0.44

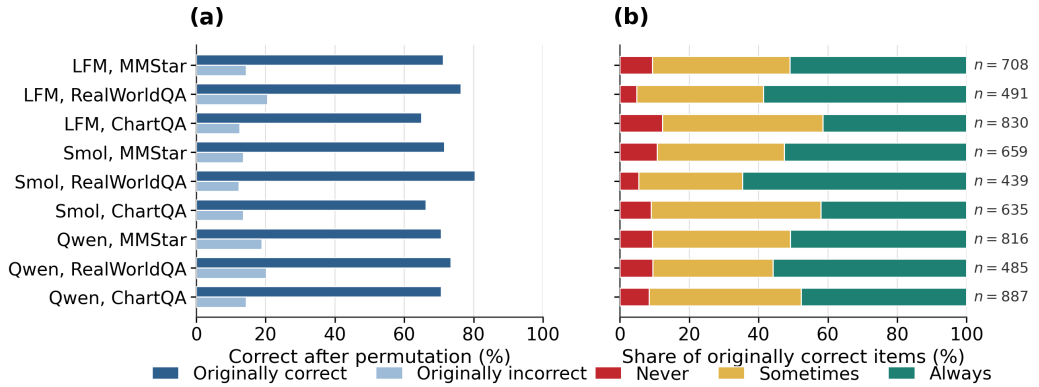


Figure 8: (a) separator-free k=2 transitions normalized within outcome strata, showing the share of originally correct trials that remain correct and the share of originally incorrect trials that become correct for each model and benchmark. (b) per-item spectrum among originally correct items (never, sometimes, or always correct across the 5 derangements), with the number of contributing items annotated beside each bar.

A.15 Retention-stratified persistence

Table 5 and Figure 9 report persistence grouped by how many directed tile-adjacency relations survive within each item's sampled derangements (zero, low, or high retention), restricted to items whose samples cover all three strata. The high-minus-zero difference is positive for all nine model-benchmark pairs, from $+3.21$ to $+16.91$ percentage points, and spans $+6.67$ to $+8.20$ percentage points on MMStar, where the three models are evaluated on the same items under the same sampled permutations. Retained adjacency covaries with tile displacement and with row and column ordering, both included in the released per-permutation records, so we read this comparison as descriptive and not causal.

A.16 Target-specific persistence gap

Table 6 and Figure 10 show that the target-specific persistence gap is positive with a 95% confidence interval that excludes zero in every one of the 27 model-benchmark-grid combinations, ranging from about $+17$ percentage points at k=8 for some pairs to over $+50$ percentage points at k=2 on ChartQA. Permuted-target correctness remains substantially above both conditional baselines on the same originally correct items.

Table 5: Retention-stratified separator-free k=2 persistence (percent), items with all three strata observed. High-minus-zero differences use Holm-adjusted two-sided p-values within the 9-contrast family.

Model	Benchmark	Zero	Low	High	High−Zero (Holm p)
LFM	MMStar	67.46	70.26	75.46	+8.00 (0.002)
LFM	RealWorldQA	72.46	76.45	79.08	+6.61 (0.005)
LFM	ChartQA	59.52	62.70	76.43	+16.91 (0.002)
Smol	MMStar	70.42	71.51	77.09	+6.67 (0.002)
Smol	RealWorldQA	80.89	81.75	84.10	+3.21 (0.123)
Smol	ChartQA	61.69	64.82	73.35	+11.66 (0.002)
Qwen	MMStar	69.38	69.46	77.58	+8.20 (0.002)
Qwen	RealWorldQA	71.16	72.89	77.38	+6.22 (0.010)
Qwen	ChartQA	68.06	67.28	83.92	+15.85 (0.002)

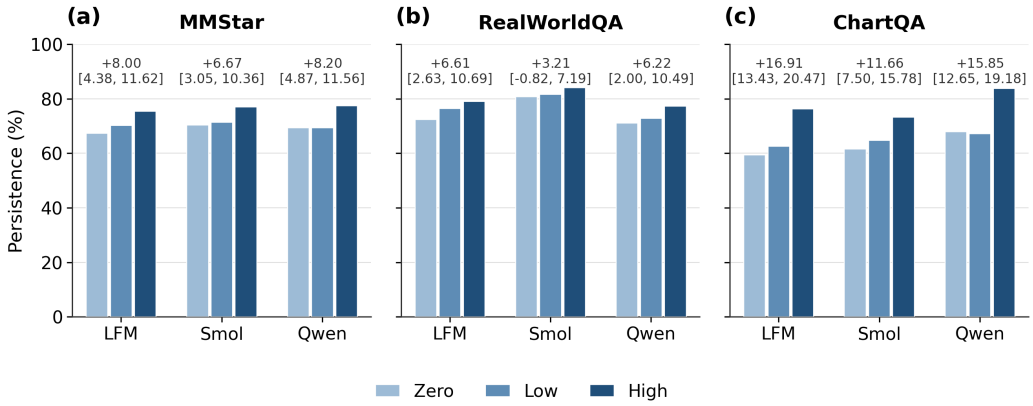


Figure 9: Retention-stratified separator-free k=2 persistence for each model and benchmark, with (a) MMStar, (b) RealWorldQA, and (c) ChartQA, grouped by retained directed adjacency (zero, low, or high) on items whose 5 samples cover all three strata. Each model group is annotated with the high-minus-zero difference and its nested design bootstrap 95% interval in pp.

A.17 Separator-line ablation

Table 7 gives the paired persistence differences summarized in Section 5. Figure 11 shows persistence for separated derangements, using the identical permutation orders as the separator-free condition in Figure 1. Because separator density grows with the grid size, we use these conditions for the paired comparisons within each grid size reported in Table 7.

A.18 Diagnostics and implementation details

No model-benchmark pair exceeded the 5% extraction-failure threshold set for exclusion, so all nine pairs are reported. Extraction failure rates on the padded-original condition stay below 2.5% everywhere; the blind condition on ChartQA reaches up to 12.3%, and extraction failure on the permuted k=2 condition never exceeds 2.6%. Invalid multiple-choice options occur at most 0.4% of the time on any MMStar pair. We used `torch` 2.11.0, `transformers` 5.9.0, `datasets` 4.4.1, `Pillow` 11.3.0, `numpy` 2.0.2, and `scipy` 1.15.3. Datasets are obtained through the Hugging Face `datasets` library, and model and dataset revisions are pinned to full commit hashes.

Table 6: Target-specific persistence gap (pp). For each originally correct item, separator-free permuted-target correctness minus the mean conditional correctness of the content-removed baselines (blind and 3 random-image maps), averaged over items.

Model	Benchmark	k=2	k=4	k=8
LFM	MMStar	+36.43	+27.33	+18.86
LFM	RealWorldQA	+32.95	+28.45	+22.54
LFM	ChartQA	+48.90	+27.85	+17.13
Smol	MMStar	+34.23	+26.77	+20.60
Smol	RealWorldQA	+26.89	+25.34	+20.79
Smol	ChartQA	+44.83	+30.59	+18.88
Qwen	MMStar	+35.59	+26.96	+18.63
Qwen	RealWorldQA	+28.48	+24.90	+18.37
Qwen	ChartQA	+54.81	+35.81	+20.18

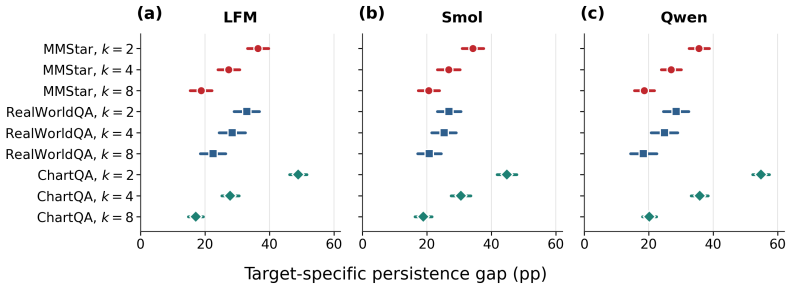


Figure 10: Target-specific persistence gap for each benchmark and grid size k, defined in Section 3, with (a) LFM, (b) Smol, and (c) Qwen. Segments are item-cluster 95% bootstrap intervals and the vertical line marks zero.

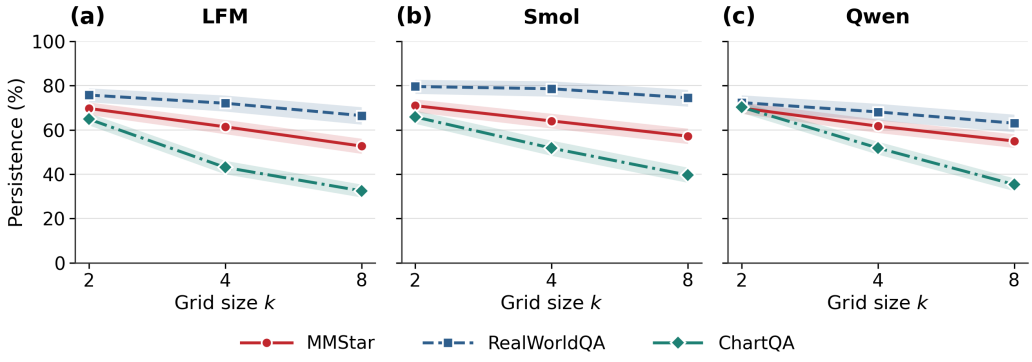


Figure 11: Persistence against grid size k for separated derangements (one-pixel channel-mean separator at each internal boundary), using the identical permutation orders as the separator-free condition, with one curve per benchmark. (a) LFM, (b) Smol, (c) Qwen. Shaded bands are two-level bootstrap 95% intervals.

Table 7: Separator-line ablation summary. Paired persistence differences (separated minus separator-free) use identical permutation orders in both conditions; Holm p is the adjusted two-sided p-value within the 27-contrast family. The four contrasts shown include the smallest adjusted p-value in the family.

Model	Benchmark	k	Persistence difference (pp)	Holm p
LFM	MMStar	2	-1.55 [-2.57, -0.59]	0.065
Qwen	RealWorldQA	4	-1.79 [-3.37, -0.21]	0.858
Smol	RealWorldQA	2	-0.77 [-1.55, -0.05]	1.00
LFM	ChartQA	8	-0.76 [-1.93, $+0.40$]	1.00