HIDDEN SENSITIVITY IN SPATIAL REASONING BENCHMARKS: DIAGNOSIS AND FACTOR-BALANCED EVALUATION

Anonymous authors
Paper under double-blind review

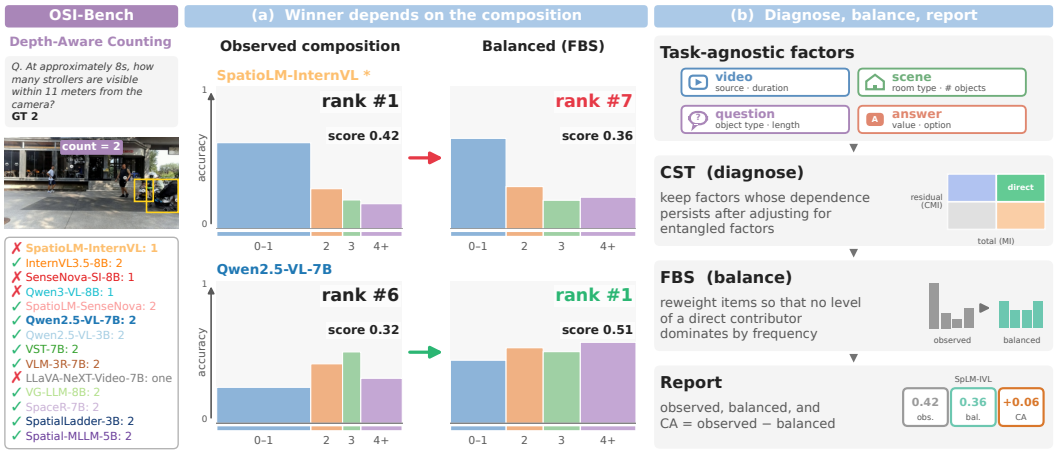

Figure 1: **Benchmark scores are composition-weighted averages. (a)** On OSI-Bench (Wu et al., 2026d) depth-aware counting, about half of the items have answers of 0 or 1. Balancing the diagnosed factors changes the rankings substantially: SpatioLM-InternVL moves from rank 1 to 7, while Qwen2.5-VL-7B moves from rank 6 to 1. **(b)** Our framework first uses CST to *diagnose* task-agnostic factors whose composition systematically affects scores, then applies FBS to *balance* their contributions. We report the observed score, balanced score, and composition advantage together to expose the effect of benchmark composition on model comparisons.

ABSTRACT

Spatial reasoning benchmarks for vision-language models (VLMs) report aggregate scores that implicitly weight evaluation conditions by their frequency in the benchmark. Across four video spatial reasoning benchmarks (VSI-Bench, ReVSI, VSTI-Bench, and OSI-Bench) and 14 VLMs, we show that the distributions of *task-agnostic factors*, such as ground-truth answer value and queried object type, are highly skewed. Models exhibit distinct performance patterns across these factors, and plausible changes in composition alone often reorder leaderboards. We quantify this effect through a model's *composition advantage*, defined as the difference between its observed score and a factor-balanced counterpart. To identify consequential factors, we introduce the Composition Sensitivity Test (CST), an information-theoretic diagnostic that retains factor-performance dependencies that persist after adjustment for entangled factors. It identifies answer value and object type as recurring sources of composition sensitivity, while also capturing additional influential factors when present. The Factor-Balanced Score (FBS) then reweights benchmark items to alleviate the dominance of frequently represented levels of diagnosed factors in the aggregate score. Across the four benchmarks, balancing changes the ranks of 53% of model-task pairs and the top-ranked model on 10 of 29 affected tasks. On object counting, for example, answers are concentrated in a narrow range of values; on VSI-Bench, balancing this distribution decreases a spatially fine-tuned model's score by 0.25, while a general-purpose model whose score barely changes rises from eighth to first place. This illustrates how benchmark compositions can unintentionally favor some models over others. Moreover, under systematic composition changes, FBS reduces score variability by 63–71% relative to the observed scores. Together, these results show that benchmark composition can notably affect conclusions about spatial reasoning performance of

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.

VLMs. We therefore advocate jointly reporting observed scores, balanced scores, and composition advantages to make this influence explicit.

1 INTRODUCTION

Spatial reasoning—counting objects, estimating sizes and distances, understanding room-scale geometry, and tracking camera motion—is a key capability for vision-language models (VLMs), particularly as they become the perception-and-reasoning core of embodied agents (Chen et al., 2025; Lee et al., 2025a; Team et al., 2025; Koo et al., 2025). Recent benchmarks such as VSI-Bench (Yang et al., 2025a), ReVSI (Zhang et al., 2026b), VSTI-Bench (Fan et al., 2026), and OSI-Bench (Wu et al., 2026d) provide curated task suites for measuring these capabilities. Their resulting leaderboards help researchers understand current model strengths and weaknesses and guide decisions about training data, model architectures, or computational resources for developing more spatially capable VLMs.

However, focusing only on final aggregate scores can overlook an important aspect that shapes the benchmark results: the benchmark's *composition*. Building a benchmark involves many practical design choices, such as which scene datasets to draw from, which object categories and room types to include, and what range of answer values to query. These choices determine the distributions of *task-agnostic factors*—properties of the evaluation data that are not themselves the spatial capabilities being assessed, yet can influence measured performance. Such choices are typically treated as implementation details, fixed at construction time and rarely revisited during evaluation. We argue that their influence should also be explicitly taken into account when interpreting what benchmark results imply about a model's spatial reasoning ability.

Our work is motivated by two empirical observations. *First*, task-agnostic factors in current spatial reasoning benchmarks are often highly skewed: certain input video sources, object categories, and answer ranges are substantially over-represented. For example, VSI-Bench's object counting task is concentrated in a narrow range of low ground-truth values, and VSTI-Bench's camera-object distance task asks about "table" in a quarter of its queries (225 of 905). *Second*, aggregate benchmark scores can be strongly coupled to these imbalanced compositions, such that reported performance gaps partly reflect which factor configurations happen to dominate the evaluation set. This is demonstrated in the examples in Fig. 2. Without accounting for this dependence, leaderboard results can conflate differences in model performance with differences induced by benchmark composition.

Manually pinpointing which task-agnostic factors matter is difficult because each benchmark contains many candidate factors that vary across tasks and query types, while their distributions are often strongly entangled with one another. Therefore, we introduce a structured benchmark-diagnosis framework that automatically identifies and quantifies hidden dependencies on these factors in spatial reasoning benchmarks. Building on this diagnosis, we further propose a factor-balanced aggregation scheme that reweights benchmark items to reduce the disproportionate influence of skewed diagnosed factors, yielding aggregate scores that are less dominated by the composition of specific factors. Specifically, our framework consists of two components:

(1) Composition Sensitivity Test (CST). CST identifies which task-agnostic factors are key contributors to aggregate benchmark scores. For each factor, CST compares its marginal dependence on per-item scores with the residual dependence that remains after adjusting for other factors. A factor is identified as a *direct contributor* when this dependence remains substantial after accounting for factor entanglement.

(2) Factor-Balanced Score (FBS). Given the identified direct contributors, FBS reweights benchmark items so that the composition of each contributor becomes approximately uniform, providing a frequency-neutral reference. We then analyze how this balancing changes model rankings and scores, revealing notable composition sensitivity in previous aggregation and showing that FBS yields more composition-robust evaluation.

We apply our diagnostic framework across four video spatial reasoning benchmarks, covering 39 tasks and 14 VLMs spanning both general-purpose and spatially fine-tuned models. Across benchmarks, CST identifies ground-truth answer value and queried object type as recurring sources of composition sensitivity, while ten relational tasks involving relative direction, distance, or position have no direct contributor. Applying FBS changes 53% of model rankings and also changes the top-1 model on 10 of the 29 tasks that had a direct contributor. On counting tasks, composition advantages reach -0.19 for general-purpose models and $+0.25$ for spatially fine-tuned models, and balancing substantially

reduces composition-induced score variability under resampling. Together, these results show that benchmark composition can significantly affect reported model comparisons, motivating the joint assessment of aggregate scores, balanced scores, and composition advantages when evaluating VLM spatial reasoning.

In summary, our contributions are as follows:

- We formalize composition sensitivity and demonstrate how skewed task-agnostic factor distributions shape model comparisons across four spatial reasoning benchmarks (Sec. 3).
- We introduce CST, an information-theoretic benchmark diagnostic that identifies influential task-agnostic factors whose associations with scores remain large even after adjusting for entangled factors (Sec. 3.2).
- We introduce FBS, a factor-balanced score through reweighting, and *composition advantage*, which quantifies each model's gain or loss from the observed composition relative to the balanced reference. Our evaluation of 14 VLMs across 39 tasks reveals substantial re-ranking and improved score stability under composition resampling (Sec. 4).

2 Related Work

2.1 Spatial Reasoning in Vision-Language Models

Improving and evaluating spatial reasoning capabilities of VLMs has emerged as a central challenge in the recent literature on embodied agents. One line of research takes a *data-centric* approach: SpatialVLM (Chen et al., 2024) showed that scalable spatial VQA data can substantially improve spatial reasoning. Another line takes an *architectural* approach by introducing more spatially grounded model designs. SpatialRGPT (Cheng et al., 2024) and VLM-3R (Fan et al., 2026) augment visual representations with geometric cues such as depth, segmentation, or 3D features, enabling spatial queries beyond raw pixel inputs. An orthogonal line of work has developed benchmarks for evaluating VLM spatial reasoning. VSI-Bench (Yang et al., 2025a) provides a comprehensive taxonomy spanning a broad range of video-based spatial reasoning tasks. More recent benchmarks extend this direction by improving visual and annotation quality (Zhang et al., 2026b), incorporating temporal axis into questions (Fan et al., 2026), and expanding from indoor scenes to outdoor environments (Wu et al., 2026d). We discuss additional recent work in App. C.

2.2 Benchmark and Evaluation Validity

In the era of foundation models with near-human-level capabilities (Comanici et al., 2025; Hurst et al., 2024; Bai et al., 2025a), faithfully assessing their abilities along specific dimensions has become more important than ever. A growing literature questions whether benchmark scores truly reflect the capabilities they claim to measure (Neuhaus et al., 2026). Singh et al. (2025) identify systematic biases in Chatbot Arena, while Alzahrani et al. (2024) show that LLM leaderboards can be perturbed by minor design choices in benchmarks such as answer-choice ordering and scoring procedures. Zhao et al. (2026) further argue that LLM chain-of-thought (CoT) gains may largely stem from in-distribution reasoning patterns rather than general reasoning ability. Moreover, Gorla & Puduppully (2026) attribute apparent gains on tabular benchmarks to memorization rather than generalization.

Similar concerns have been raised in visual and spatial reasoning. VQA-CP (Agrawal et al., 2018) exposes reliance on language priors, while MIRAGE (Asadi et al., 2026) shows that strong multimodal benchmark performance can persist even without image inputs. VPBench (Feng et al., 2025) demonstrates sensitivity to visual marker design and inference settings, and Ma et al. (2026) show that text-only models can match or surpass 3D-LLMs on 3D reasoning benchmarks. For spatial reasoning, Brown et al. (2025b) diagnose non-visual shortcuts in VSI-Bench and prune susceptible examples, while ReVSI (Zhang et al., 2026b) addresses annotation errors and questions that are unanswerable from the available video frames. Our work complements these efforts by introducing a benchmark diagnostic framework that identifies task-agnostic factors whose associations with benchmark scores remain after disentanglement, together with a factor-balanced metric that reduces the dominance of their skewed compositions in aggregate scores.

3 Composition Sensitivity in Spatial Reasoning Benchmarks

We study how benchmark composition influences aggregate spatial reasoning scores and identify which factors are responsible for this sensitivity. In Sec. 3.1, we view each benchmark as a com-

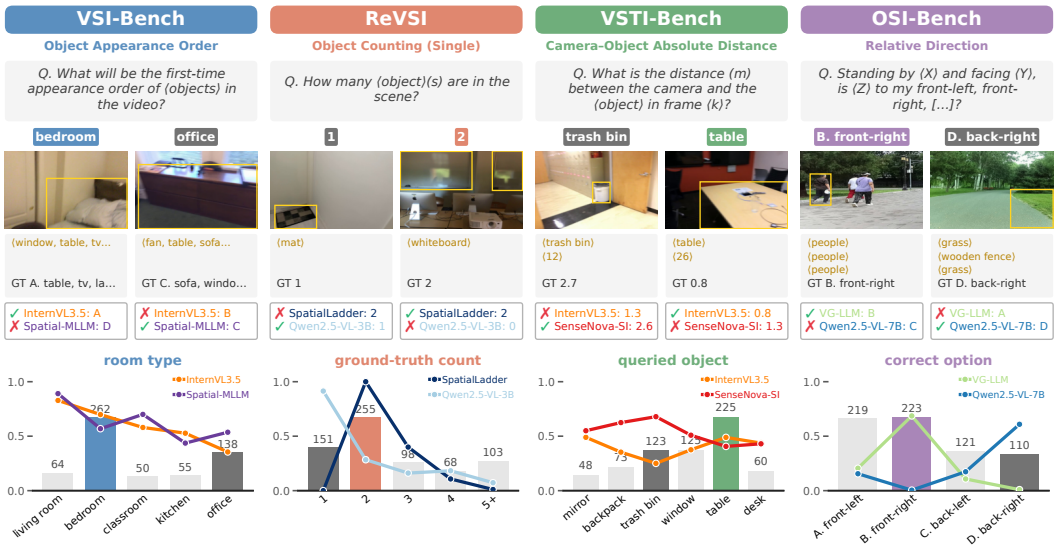

Figure 2: **Composition sensitivity across four spatial reasoning benchmarks.** For each benchmark, we show one task, one task-agnostic factor, and two models whose performance varies differently across factor levels. The factor distributions are skewed, with a single level accounting for 25–42% of the task items. Because the models' relative performance changes across levels, the benchmark composition can determine which model ranks higher.

position of *task-agnostic factors* (*e.g.*, properties of the video, scene, question, and answer) and formalize the aggregate score as a composition-weighted average of model performance across the benchmark. We then show that these factor distributions are often skewed and that models exhibit different performance across factor levels, making aggregate scores and leaderboard rankings sensitive to benchmark composition. In Sec. 3.2, we ask which task-agnostic factors most strongly shape benchmark scores and leaderboard rankings through their composition. We introduce an information-theoretic diagnostic to identify such influential factors. These diagnosed factors then form the basis of the factor-balanced evaluation in the following section (Sec. 4).

3.1 Benchmark is a Composition of Task-Agnostic Factors

We take a closer look at an often-overlooked aspect of benchmark construction: how benchmark items are distributed along factors that are not themselves the target of the spatial task. We characterize benchmark items through measurable *task-agnostic factors*: properties of the input video (*e.g.*, source dataset, duration, and brightness), the scene (*e.g.*, room type and number of objects), the question (*e.g.*, queried object type), and the answer (*e.g.*, answer value or option). Representative factors are illustrated in Fig. 2, with the full catalogue provided in App. D.

For a fixed task $t \in \mathcal{T}$, we suppress the task index for clarity. Let $\mathcal{I}$ denote its benchmark items and $s_{m,i}$ the score of model m on item i. Each item is associated with a configuration of task-agnostic factor values. Let $\mathcal{C}$ denote the set of configurations represented in the benchmark and $\mathcal{I}_c \subseteq \mathcal{I}$ the items belonging to configuration $c \in \mathcal{C}$. We define

$$p_{\text{bench}}(c) = \frac{|\mathcal{I}_c|}{|\mathcal{I}|}, \qquad \bar{s}_m(c) = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} s_{m,i}, \tag{1}$$

where $p_{\text{bench}}(c)$ is the empirical frequency of the configuration c and $\bar{s}_m(c)$ is model m's mean score on that configuration. The observed aggregate score is then

$$S_m^{\text{obs}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_{m,i} = \sum_{c \in \mathcal{C}} p_{\text{bench}}(c)\, \bar{s}_m(c). \tag{2}$$

We refer to S_m^{obs} as the *observed score*. Thus, the aggregate depends on both how a model performs across factor configurations, $\bar{s}_m(c)$, and how frequently those configurations occur, $p_{\text{bench}}(c)$. Construction choices—which videos to include, which objects to query, and which answer ranges to cover—implicitly determine how much each configuration contributes to the final score. When models perform differently across configurations, this composition can affect not only their aggregate scores but also their relative rankings.

3.2 Which Factors Drive Composition Sensitivity?

Problem definition. Our main question is whether certain task-agnostic factors critically shape aggregate benchmark scores through their composition, and if so, which factors matter for each spatial task. This identification is non-trivial because factor distributions are mutually entangled: an apparent association between one factor and model performance may instead be inherited from another correlated factor. We therefore seek factors whose associations with the score persist after adjustment for other factors. For each task t, let Y_t denote the per-item score averaged across models and discretized into ten equal-width bins, and let each factor f be represented by a categorical variable X_f with finite level set $\mathcal{V}_f$. For simplicity, we use Y in the remainder of this section.

Mutual information and measuring dependence. To quantify the dependence between task-agnostic factors and benchmark scores, we use mutual information (MI) (Shannon, 1948; Cover et al., 1991), which captures statistical dependence without assuming a particular functional relationship. For a *focal* factor f and an *adjustment* factor $g \neq f$, we define

$$I_f = I(Y; X_f) = \mathbb{E}\Big[\log \tfrac{p(Y,X_f)}{p(Y)p(X_f)}\Big], \quad I_{f|g} = I(Y; X_f \mid X_g) = \mathbb{E}\Big[\log \tfrac{p(Y,X_f|X_g)}{p(Y|X_g)\,p(X_f|X_g)}\Big]. \tag{3}$$

The marginal MI I_f measures the *total dependence* between factor X_f and the score, including dependence shared with entangled factors. The conditional MI (CMI) $I_{f|g}$ measures the *residual dependence* that remains after conditioning on X_g. At the population level, $I_{f|g} = 0$ if and only if $Y \perp X_f \mid X_g$. Comparing I_f and $I_{f|g}$ therefore indicates whether the factor-score association persists after adjustment for another factor. For each task t, we retain the ten factors with the highest marginal MI with Y_t as the candidate set $\mathcal{F}_t$ (App. E).

Composition Sensitivity Test (CST). We introduce CST to answer the following question: how much of a focal factor's association with the score persists after adjustment for other factors? For task t and each focal factor f, we average its residual dependence over single-factor adjustments and define the *sensitivity ratio* as

$$\bar{I}_{f|g} = \frac{1}{|\mathcal{F}_t| - 1} \sum_{g \in \mathcal{F}_t \setminus \{f\}} I_{f|g}, \qquad \mathcal{R}_f = \frac{\bar{I}_{f|g}}{I_f}, \tag{4}$$

with $\mathcal{R}_f := 0$ when $I_f = 0$. For $I_f > 0$, this ratio measures average residual dependence relative to total dependence: values near one indicate persistence, values below one indicate attenuation, and values above one indicate stronger dependence after adjustment. Two further criteria turn the ratio into a classification. First, proximity to one is not a decision rule: the ratio of a persistent factor rarely equals one exactly. We therefore fix a band $1/r \leq \mathcal{R}_f \leq r$ with $r = 1.5$ and treat a factor as persistent when its dependence moves by less than a factor of r after adjustment. Second, the ratio alone does not establish whether the dependence is substantial: two negligible quantities can also yield a ratio near one. We therefore pair it with a magnitude threshold $\gamma_t = cH(Y_t)$, scaled by the score-bin entropy of each task t, with $c = 0.02$. Both constants are fixed across tasks, with the sensitivity to c evaluated in App. E. Together, these criteria classify each factor as a *direct contributor*, *suppressed*, *surface correlate*, or *negligible*, as specified in Tab. 1. Here, "direct" denotes substantial residual dependence on average across adjustments, not a causal effect. The resulting direct contributor set $\mathcal{G}_t \subseteq \mathcal{F}_t$ specifies the factors balanced in Sec. 4.

Table 1: **CST sensitivity regimes.** $\gamma_t = cH(Y_t)$ with $c = 0.02$; $\mathcal{R}_f = \bar{I}_{f|g}/I_f$ with band constant $r = 1.5$.

Regime	$I_f > \gamma_t$ and $\mathcal{R}_f \leq r$	otherwise
$\bar{I}_{f\|g} > \gamma_t$ and $\mathcal{R}_f \geq 1/r$	Direct	Suppressed
otherwise	Surface	Negligible

CST diagnosis across four benchmarks. We apply CST across four video spatial reasoning benchmarks to identify task-agnostic factors that exhibit high composition sensitivity. The top row of Fig. 3 displays one representative task from each benchmark, illustrating the four CST regimes (full results in App. E). *Direct* (VSI-Bench's counting): answer value and object type have $I_f/H(Y_t) = 0.26$ and 0.16, and retain 94% and 90% of their dependence after adjustment, while all other factors fall below the threshold γ_t. *Negligible* (ReVSI's relative direction, backward-hard): object type is the only direct contributor, file size is suppressed, while the remaining factors are negligible. *Suppressed* (VSTI-Bench's relative position, up/down): several factors show increase

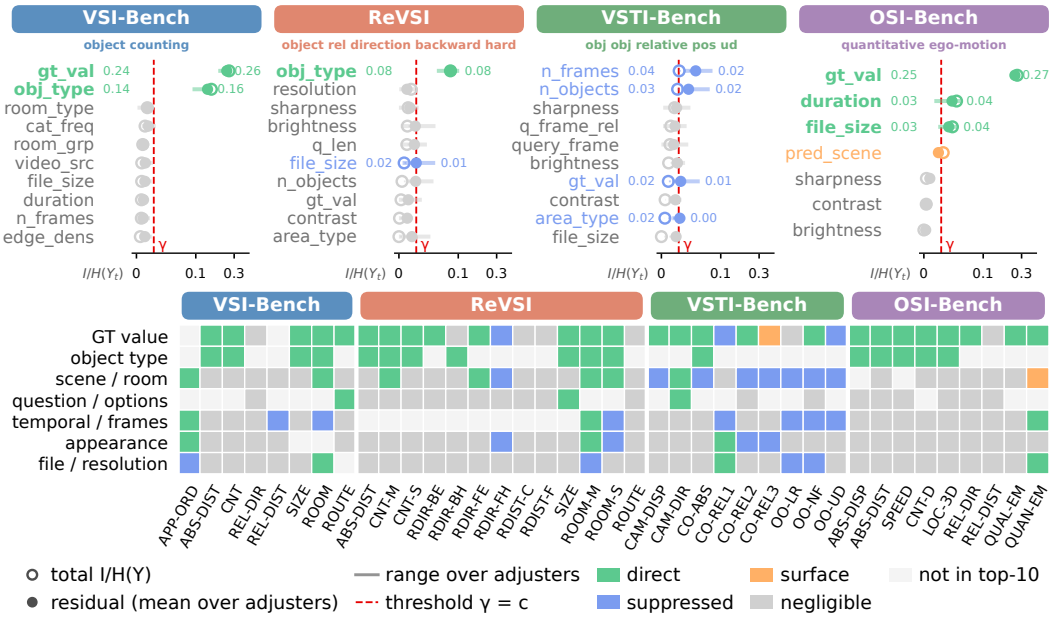

Figure 3: **CST diagnosis.** Top: total and residual dependence of the candidate factors on one task per benchmark ($c = 0.02$; asinh axis); the residual $\bar{I}_{f|g}/H(Y_t)$ is printed left of the axis and the total $I_f/H(Y_t)$ right of the markers. Bottom: regime of each factor family's strongest candidate on all 39 tasks; columns without a direct contributor are mostly relational tasks. Task abbreviations are listed in Tab. B.2.

in dependence after adjustment, but none qualifies as a direct contributor. *Surface correlate* (OSI-Bench's quantitative ego-motion): scene type falls below the threshold after adjustment, while answer value, duration, and file size remain direct contributors.

The bottom panel of Fig. 3 summarizes all 39 tasks by factor family. *Two factors dominate:* answer value is a direct contributor on 26 of the 39 tasks and object type on 17. The ten tasks with no direct contributor are almost all relational (relative direction, distance, or position); FBS has nothing to balance on these tasks, so they keep their observed scores, and the balancing in Sec. 4 applies to the remaining 29.

4 Towards Factor-Balanced Evaluation of Spatial Reasoning

Building on the CST diagnosis, we introduce the Factor-Balanced Score (FBS), which reweights benchmark items so that the diagnosed factors contribute more evenly to the aggregate score (Sec. 4.1). We evaluate FBS from three complementary perspectives: (1) *re-ranking*, which captures changes in model leaderboards, (2) *composition advantage*, which measures how much each model benefits from or is penalized by the specific benchmark composition, and (3) *robustness under composition resampling*, which tests whether factor-balanced comparisons are less sensitive to changes in composition. We then apply FBS across all four benchmarks and examine how factor balancing changes model comparisons in practice (Sec. 4.3).

4.1 Factor-Balanced Score (FBS)

Recall from Eq. (2) that the observed score S_m^{obs} implicitly weights factor configurations according to their frequency in the benchmark. For task t, let $\mathcal{G}_t$ denote the direct contributors identified by CST. We seek an alternative aggregation of the same benchmark items in which the observed levels of these diagnosed factors contribute more evenly. We assign each item $i \in \mathcal{I}$ a nonnegative weight w_i, shared across all models and normalized such that $\sum_{i\in\mathcal{I}} w_i = 1$. These weights are chosen to approximately balance every diagnosed factor $f \in \mathcal{G}_t$ across its observed levels $v \in \mathcal{V}_f$. That is, after reweighting, each diagnosed factor should follow an approximately *uniform* marginal distribution. The resulting Factor-Balanced Score (FBS) is

$$S_m^{\text{bal}} = \sum_{i\in\mathcal{I}} w_i \, s_{m,i}. \tag{5}$$

Thus, S_m^{obs} and S_m^{bal} evaluate the same model predictions on the same benchmark items but differ in their aggregation weights. The observed score follows the benchmark frequencies, whereas FBS

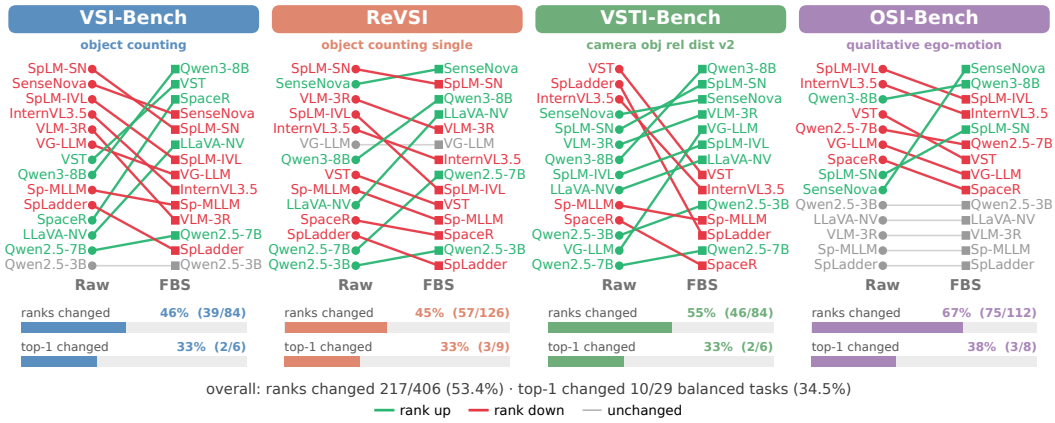

Figure 4: **Re-ranking under factor balancing. Top:** For one task from each benchmark whose top-ranked model changes, we compare model ranks under the observed score and FBS. Green indicates rank increases, red rank decreases, and gray unchanged ranks. **Bottom:** For each benchmark, we report the fraction of model-task ranks that change and the fraction of balanced tasks whose top-ranked model changes. Overall, 217 of 406 model-task ranks (53.4%) change, and the top-ranked model changes on 10 of 29 balanced tasks (34.5%).

reduces the dominance of frequently represented levels of CST-diagnosed factors through *reweighting*. Importantly, the weights depend only on benchmark composition and are shared across all evaluated models.

We obtain w_i using stable balancing weights (SBW) (Zubizarreta, 2015), which approximately satisfy the uniform balancing targets while limiting weight dispersion. We provide feasibility details and ablations of alternative reweighting schemes in App. F, along with SBW's tolerance threshold δ. When CST identifies no direct contributor for a task, we apply no reweighting and set $S_m^{\text{bal}} = S_m^{\text{obs}}$.

4.2 Setup

Analysis targets. We analyze the effect of FBS along three complementary axes: how balancing changes model rankings, which models benefit from the benchmark composition, and whether the resulting scores are more robust to composition changes.

- *Re-ranking.* We compare the rankings induced by S_m^{obs} and S_m^{bal} using rank changes, top-1 changes, and Kendall's τ (Kendall et al., 1938). This measures how strongly benchmark composition affects the leaderboard conclusions of the original evaluation.
- *Composition advantage.* We define $\Delta S_m = S_m^{\text{obs}} - S_m^{\text{bal}}$ to quantify how much model m benefits from or is disadvantaged by the observed composition relative to the balanced one. Comparing ΔS_m across models within a task shows whether composition widens, narrows, or reverses observed performance gaps.
- *Robustness to composition changes.* Finally, we test whether factor balancing makes model comparisons more robust to composition shifts. For each task, we vary the benchmark composition along its strongest direct contributor and compare the ranking and score stability of the observed and balanced evaluations.

Models and Benchmarks. We evaluate 14 VLMs across four video spatial reasoning benchmarks: VSI-Bench (Yang et al., 2025a), ReVSI (Zhang et al., 2026b), VSTI-Bench (Fan et al., 2026), and OSI-Bench (Wu et al., 2026d). The model set includes both general-purpose VLMs and spatially fine-tuned variants. The full list is provided in App. B. CST identifies direct contributors for 29 of the 39 tasks; for each such task, we report the observed score, balanced score, and composition advantage for every model, with full results in App. G.

4.3 Key Analysis and Takeaways

Factor balancing substantially perturbs leaderboards. Fig. 4 summarizes the effect of factor balancing on model rankings. Across the 29 balanced tasks, 217 of 406 model-task ranks (53.4%) change, and the top-ranked model changes on 10 of 29 tasks (34.5%). This effect is observed consistently across benchmarks: the fraction of changed ranks ranges from 45.2% on ReVSI to 67.0% on OSI-Bench, while top-1 flips occur on 2/6 VSI-Bench tasks, 3/9 ReVSI tasks, 2/6 VSTI-Bench tasks, and 3/8 OSI-Bench tasks.

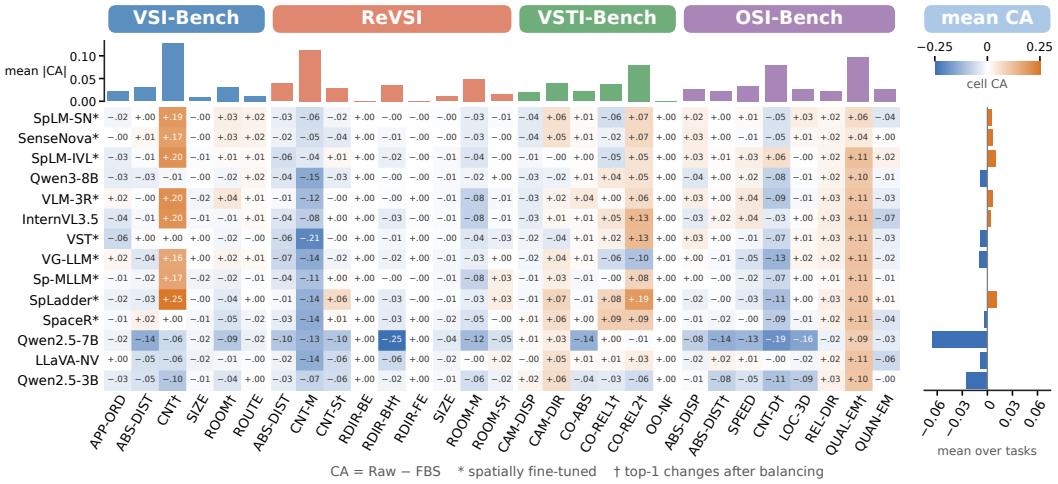

Figure 5: **Composition advantage of every model on every balanced task.** Cells: $\Delta S_m = S_m^{\text{obs}} - S_m^{\text{bal}}$ (CA in the figure); orange means the observed composition inflates the model's score, blue that it deflates it. Top: mean $|\Delta S_m|$ per task; right: mean ΔS_m per model. The two all-zero columns are tasks whose observed levels already lie within δ of the balanced target. Task abbreviations are listed in Tab. B.2.

The representative examples in Fig. 4 show that these changes can be substantial rather than local swaps. On VSI-Bench's object counting, SpatioLM-SenseNova falls from first to fifth, while Qwen3-VL-8B rises from eighth to first. On ReVSI's single-object counting, balancing also changes the winner while substantially reordering the middle of the leaderboard. On VSTI-Bench's camera-object relative distance v2, VST-7B drops from first to eighth while Qwen3-VL-8B rises from seventh to first. Similarly, on OSI-Bench's qualitative ego-motion, the top-ranked model changes after balancing. These examples illustrate that changing only the aggregation weights of the same benchmark items can materially alter the model comparisons supported by the leaderboard.

Spatially fine-tuned models often fall below their base models. Factor balancing can also change conclusions within model families, reversing the relative ordering between spatially fine-tuned models and their backbone counterparts. On VSI-Bench's object counting, SpatioLM-SenseNova falls from first to fifth while its SenseNova-SI-8B backbone moves from second to fourth. Similar reversals occur on ReVSI's single-object counting and VSTI-Bench's camera-object relative distance v2, where SpatialLadder-3B and SpaceR-7B fall below their respective Qwen2.5-VL backbones. On OSI-Bench, VST-7B moves below Qwen2.5-VL-7B on qualitative ego-motion, and SpatioLM-InternVL falls below InternVL3.5-8B on depth-aware counting. These reversals show that the apparent advantage of spatial specialization over its underlying backbone can itself depend on benchmark composition: the models and predictions remain unchanged, and only the aggregation weights differ.

Composition advantage is pronounced on counting tasks. The heatmap in Fig. 5 shows the task-wise variation of ΔS_m. Notably, it reveals that counting tasks show particularly large magnitudes across models: mean $|\Delta S_m|$ reaches 0.127 on VSI-Bench's counting, 0.113 on ReVSI's multi-object counting, and 0.080 on OSI-Bench's depth-aware counting. On counting tasks, answer value and object type recur as direct contributors under CST, highlighting the influence of their benchmark distributions. On VSI-Bench's counting, seven of nine spatially fine-tuned models have a positive composition advantage of $\Delta S_m \in [0.16, 0.25]$, with InternVL3.5-8B also at 0.20, whereas Qwen2.5-VL-3B/7B and LLaVA-NeXT-Video have negative ΔS_m. This difference substantially widens observed performance gaps: after balancing, SpatialLadder's lead over Qwen2.5-VL-3B shrinks from 0.38 to 0.03, and VG-LLM's lead over Qwen2.5-VL-7B from 0.29 to 0.07. By contrast, all models have negative ΔS_m on ReVSI's multi-object counting and positive ΔS_m on OSI-Bench's qualitative ego-motion, indicating that benchmark composition can also shift all models in the same direction.

Composition advantage is task-specific, rather than a model property. The same model can benefit from benchmark composition on one task but be disadvantaged on another. For example, SpatialLadder has $\Delta S_m = 0.25$ on VSI-Bench's counting and 0.19 on VSTI-Bench's relative distance v2, but -0.14 on ReVSI's multi-object counting. Similarly, VST-7B is nearly neutral on VSI-Bench's counting, but has $\Delta S_m = -0.21$ on ReVSI's multi-object counting, and 0.13 on VSTI-Bench's relative distance v2. Qwen2.5-VL-7B is a notable exception, showing the most negative ΔS_m on 16

of 29 tasks, with a mean of -0.066. We therefore report ΔS_m at the task level and caution against interpreting its average as an intrinsic property of the model.

Factor balancing improves robustness to composition changes. Finally, we test robustness by systematically perturbing the composition of a task-agnostic factor. For each task, we select the top-1 direct contributor (*i.e.*, the factor with the highest marginal MI), and sample 200 compositions by varying how each item's sampling weight is derived along that factor. Each composition contains $n/2$ sampled items, where n is the total number of items in the task. Specifically, items are ranked by that factor's value and weighted using a Beta density with mean $0.5 + 0.35\lambda$ and concentration 2, where λ is swept over $[-1, 1]$. We obtain such subsets for all 29 tasks with at least one direct contributor. We define *ranking stability* as the mean Kendall's τ over 3,000 randomly sampled pairs of compositions, computed separately for the observed and the balanced scores, and *score stability* as $1 - \mathrm{SD}/\mathrm{SD}_{\max}$, where SD is the mean model-wise score variation across compositions and $\mathrm{SD}_{\max} = 0.200$ is the largest variation observed across all models, tasks, subset sizes, and scoring methods. Full protocol details are provided in App. G.

Fig. 6 summarizes the results. Mean ranking stability increases from 0.78 to 0.85 on VSI-Bench, 0.71 to 0.85 on ReVSI, 0.52 to 0.78 on VSTI-Bench, and 0.55 to 0.85 on OSI-Bench. Across benchmarks, FBS reduces the standard deviation of model scores across these systematically varied compositions by 63–71%. Overall, FBS consistently improves both ranking and score stability across different compositions of the critical task-agnostic factors. However, the robustness gains are not uniform across all tasks. On ReVSI's multi-object counting, ranking stability decreases slightly from 0.83 to 0.82, despite score stability rising from 0.63 to 0.89. On VSI-Bench's counting, ranking stability remains approximately 0.65, while score stability rises from 0.46 to 0.77. Overall, balancing substantially reduces composition sensitivity, although improvements in ranking stability are not guaranteed for every task.

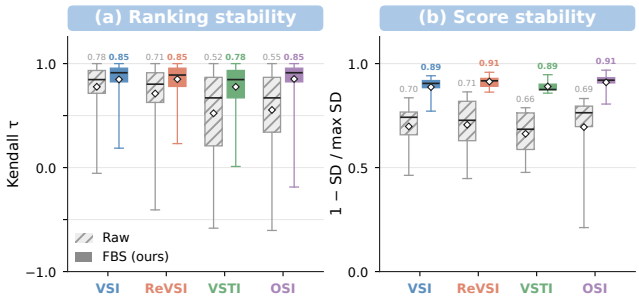

Figure 6: **Robustness to composition changes.** (a) Ranking stability: Kendall's τ between composition-wise leaderboards, averaged over tasks within each benchmark. (b) Score stability: $1 - \mathrm{SD}/\mathrm{SD}_{\max}$, shown across tasks. Boxes indicate interquartile ranges, whiskers the full range, and diamonds the means. Higher values indicate greater stability.

5 Conclusion

We revisited an often-overlooked aspect of spatial reasoning evaluation: how the composition of task-agnostic factors shapes aggregate model comparisons. Across four benchmarks, 39 tasks, and 14 VLMs, we found that factors such as ground-truth answer value and queried object type can be highly skewed and strongly associated with measured performance. We introduced CST to diagnose factor-score dependencies that persist after adjustment for entangled factors, and FBS to construct a factor-balanced evaluation. This balancing changes 53.4% of model rankings and the winner on 10 of 29 tasks, while reducing composition-induced score variability by 63–71% under composition resampling. Together, these results show that benchmark composition can significantly affect reported performance gaps and model comparisons. More broadly, we hope this framework makes the hidden influence of task-agnostic factor composition explicit, enabling more careful interpretation of open benchmarks and more deliberately balanced evaluation that provides clearer feedback for improving future spatially capable VLMs.

Limitations and future work. FBS mitigates composition imbalance in existing benchmarks through post-hoc reweighting. A more fundamental direction is to incorporate composition diagnostics during benchmark construction, enabling influential task-agnostic factors to be balanced by design. Moreover, because the *true* real-world factor distributions are unknown, we use approximately uniform marginals as a reference rather than a universally correct target. Future work could explore task-specific target distributions for FBS. We provide further discussions in App. A.

AI USE STATEMENT

In this work, generative AI tools were used to assist with editing the manuscript for clarity, readability, and presentation, and with developing and refining code for experimental plots and visualizations. All AI-assisted outputs were reviewed by the authors, and visualization code and resulting figures were checked against the underlying experimental results. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies the validity and composition sensitivity of existing spatial reasoning benchmarks and does not involve human-subject experiments or the collection of personally identifiable information. We use existing benchmarks and models for research purposes and follow their respective usage conditions. We do not identify additional ethical risks beyond those associated with the underlying datasets and models.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary to reproduce our analyses throughout the paper and appendix. In particular, we specify the CST procedure and its hyperparameters, the factor definitions and discretization rules, the construction of the factor-balanced score, and the composition-resampling protocol. The appendix additionally reports per-task results, sensitivity analyses, and ablations over the main design choices. These details are intended to support reproduction of the reported diagnoses, balanced scores, re-ranking analyses, and robustness experiments.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, et al. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *arXiv preprint arXiv:2507.08306*, 2025.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *ACL*, 2024.

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, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS*, 2021.

Eli Ben-Michael, Avi Feller, and Erin Hartman. Multilevel calibration weighting for survey data. *Political Analysis*, 32(1):65–83, 2024.

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. Sims-v: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025a.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should" train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025b.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *CVPR*, 2026.

Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

Hanyang Chen, Mark Zhao, Rui Yang, Qinwei Ma, Ke Yang, Jiarui Yao, Kangrui Wang, Hao Bai, Zhenhailong Wang, Rui Pan, et al. Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning. *arXiv preprint arXiv:2510.12693*, 2025.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. In *CVPR*, 2026a.

Yixin Chen, Yaowei Zhang, Huangyue Yu, Junchao He, Yan Wang, Jiangyong Huang, Hongyu Shen, Junfeng Ni, Shaofei Wang, Baoxiong Jia, et al. Lifting unlabeled internet-level data for 3d scene understanding. *arXiv preprint arXiv:2604.01907*, 2026b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *CVPR*, 2026c.

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 *NeurIPS*, 2024.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, et al. Spatialclaw: Rethinking action interface for agentic spatial reasoning. *arXiv preprint arXiv:2606.13673*, 2026.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language Processing (EMNLP-IJCNLP)*, pp. 4069–4082, 2019.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *ICCV*, 2025.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Jinhui Yin, Qi Wei, Tianshuo Yang, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. In *ICLR*, 2026.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *CVPR*, 2026.

Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. Visually prompted benchmarks are surprisingly fragile. *arXiv preprint arXiv:2512.17875*, 2025.

Aditya Gorla and Ratish Puduppully. The illusion of generalization in tabular language models. *arXiv preprint arXiv:2602.04031*, 2026.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

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

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. In *EMNLP*, 2025.

Juil Koo, Daehyeon Choi, Sangwoo Youn, Phillip Y Lee, and Minhyuk Sung. Toward ambulatory vision: Learning visually-grounded active view selection. *arXiv preprint arXiv:2512.13250*, 2025.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025a.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *ICCV*, 2025b.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *ICLR*, 2026.

Shijie Lian, Changti Wu, Laurence Tianruo Yang, Hang Yuan, Bin Yu, Lei Zhang, and Kai Chen. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. In *CVPR*, 2026.

Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3292–3310, 2022.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. In *NeurIPS*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Jianhui Liu, Haoze Sun, Wenbo Li, Yanbing Zhang, Rui Yang, Zhiliang Zhu, Yijun Yang, Shenghe Zheng, Nan Jiang, Jiaxiu Jiang, et al. Openspatial: A principled data engine for empowering spatial intelligence. *arXiv preprint arXiv:2604.07296*, 2026.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In *NeurIPS*, 2024.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *ICLR*, 2026.

Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In *NeurIPS*, 2022.

George Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 1955.

Yannic Neuhaus, Nicolas Flammarion, Matthias Hein, and Francesco Croce. On the out-of-distribution generalization of reasoning in multimodal llms for simple visual planning tasks. *arXiv preprint arXiv:2602.15460*, 2026.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *CVPR*, 2021.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel Dsouza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. In *NeurIPS*, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *CVPR*, 2025.

Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *ICCV*, 2019.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *NeurIPS*, 2026a.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *CVPR*, 2026b.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. Spatiolm: Towards general physical spatial intelligence in vision-language models. In *International Conference on Machine Learning (ICML)*, 2026c.

Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in mllms. In *CVPR*, 2026d.

Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. Causal-halbench: Uncovering lvlms object hallucinations through causal intervention. In *AAAI*, 2026.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025a.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025b.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. In *ECCV*, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *ICLR*, 2026b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026c.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Z Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? *arXiv preprint arXiv:2506.18322*, 2025c.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *ICLR*, 2026.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *NeurIPS*, 2025.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *arXiv preprint arXiv:2603.27437*, 2026a.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. In *ICML*, 2026b.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. In *ACL*, 2026.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *NeurIPS*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

Yang Zhou, Zixuan Huang, Sunzhu Li, Zhuo Yang, Chen Zhang, Shunian Chen, Caijun Yan, Jianyao Xu, Shunyu Liu, Weijie Fu, et al. Spatialcli: Learning to reason with spatial tools, then without them. *arXiv preprint arXiv:2607.27703*, 2026.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

APPENDIX

The appendix provides additional discussion, experimental details, and analyses supporting the main paper. Sec. A discusses the scope and interpretation of FBS; Sec. B details the evaluated models, benchmarks, tasks, and scoring; Sec. C provides additional related work; Sec. D describes the construction and preprocessing of task-agnostic factors; Sec. E and Sec. F provide additional details and ablations for CST and FBS, respectively; and Sec. G presents the full per-task results and additional analyses.

A DISCUSSION

How does FBS relate to item-level benchmark quality? Our focus on composition sensitivity is complementary to, but distinct from, prior work on item-level benchmark quality and debiasing such as Brown et al. (2025b) and ReVSI (Zhang et al., 2026b). Even after ReVSI improves annotation quality and removes flawed or unanswerable items from VSI-Bench (Yang et al., 2025a), our CST analysis in Sec. 3.2 still reveals substantial sensitivity to the composition of task-agnostic factors. This suggests that reliable spatial reasoning evaluation requires careful consideration of both item-level validity and benchmark-level composition. Our work addresses the latter by providing a systematic diagnostic for identifying the factors that drive composition sensitivity and a corresponding factor-balanced evaluation.

Does FBS provide universally more faithful model comparisons? We do *not* claim that FBS universally yields a more faithful ranking of spatial reasoning ability. The original aggregate score remains a valuable summary of performance under the benchmark's original composition. Instead, we propose FBS as a complementary evaluation component that reveals when apparent model advantages depend on which conditions dominate the benchmark. For example, on VSI-Bench's object counting, balancing reduces SpatialLadder's lead over its Qwen2.5-VL-3B backbone from 0.38 to 0.03 (Fig. 5). On VSTI-Bench's camera-object relative distance v2, SpaceR-7B falls below its Qwen2.5-VL-7B backbone after balancing, reversing their observed ordering (Fig. 4). These results show that a large aggregate gain does not always translate to consistent improvements across different evaluation configurations. Our framework makes such cases explicit, helping identify where a model's advantage weakens or reverses. We therefore recommend reporting observed scores, balanced scores, and composition advantages ΔS_m together, rather than simply replacing the original leaderboard.

Why use a uniform reference for FBS? Recall from Sec. 4.1 that FBS reweights benchmark items so that the observed levels of each diagnosed factor follow an approximately uniform marginal distribution when computing S_m^{bal}. We do *not* interpret this uniform target as a universal "real-world" distribution. Such a distribution is difficult to define, since spatial quantities and query conditions vary substantially across environments, datasets, and deployment settings. For example, Tab. A.1 shows that even widely used datasets differ considerably in their distributions of object size and object count. Mean object size ranges from 13 cm to 390 cm across the examples considered, while mean object counts range from roughly 1–2 objects in several indoor scan datasets to 56 in FSC-147. Thus, choosing any single observed dataset distribution as the reference would itself privilege a particular environment or data-collection process.

Table A.1: **Spatial-factor distributions vary substantially across datasets.** Representative mean object sizes and object counts from different datasets and benchmarks.

Object size		Object count	
Source	**Mean (cm)**	**Source**	**Mean**
YCB object set (Calli et al., 2017)	13	Indoor scan datasets[†]	1.3–2.1
MultiScan (Mao et al., 2022)	74	CV-Bench (ADE20K) (Zhou et al., 2017)	1.8
ScanNet++ (Yeshwanth et al., 2023)	80	CV-Bench (COCO) (Lin et al., 2014)	2.1
ARKitScenes (Baruch et al., 2021)	87	SPBench (ScanNet) (Li et al., 2026)	2.1
3RScan (Wald et al., 2019)	96	VSI-Bench (Yang et al., 2025a)	3.9
ScanNet (Dai et al., 2017)	98	FSC-147 (Ranjan et al., 2021)	56
SPBench (ScanNet) (Li et al., 2026)	105		
VSI-Bench (Yang et al., 2025a)	117		
KITTI (Liao et al., 2022)	390		

[†] MultiScan, ARKitScenes, ScanNet, ScanNet++, and 3RScan.

Importantly, the raw benchmark composition is not an empirically validated distribution either: its frequencies over answer values, object categories, and other factors arise from benchmark construction

choices, so retaining them implicitly privileges one particular collection outcome. In the absence of a justified external target, we therefore use approximately uniform marginals over the *observed* levels of CST-based direct contributors as a transparent, frequency-neutral reference. This factor balancing remains within the support of the benchmark rather than extrapolating to unseen conditions. When an task-specific target distribution is available, the same framework can instead balance toward that target.

Why use post-hoc reweighting? Ideally, influential task-agnostic factors should be balanced *during the benchmark construction phase*. However, this requires first identifying which factors meaningfully affect the aggregate benchmark score. Because many candidate factors can vary across spatial tasks, CST provides this diagnosis and can guide which factors future benchmarks should balance during data collection.

Moreover, re-collecting a balanced benchmark would require new data collection and re-evaluating every model, whereas FBS re-aggregate existing per-item predictions under an explicit target composition without additional model inference, as demonstrated in Sec. 4.3. Comparing the observed and balanced scores can then reveal model-specific weaknesses or robustness that are obscured by the raw aggregate alone. We therefore view post-hoc balancing and balanced benchmark construction as complementary: CST and FBS enable composition-aware auditing of existing benchmarks while also informing the design of more deliberately balanced benchmarks in the future.

B Experiment Details

This section supplements Sec. 4.2 and provides additional details on the VLM models, benchmarks and tasks used in our analysis.

Models. We evaluate our framework on 14 open-source VLMs spanning both general-purpose VLMs and spatially fine-tuned variants. Refer to Tab. B.1 for the full list. General-purpose VLMs include Qwen2.5-VL-3B/7B (Bai et al., 2025b), LLaVA-NeXT-Video-7B (Zhang et al., 2024), InternVL3.5-8B (Wang et al., 2025), and Qwen3-VL-8B (Bai et al., 2025a). Spatially fine-tuned VLMs are SpatialLadder-3B (Li et al., 2026), Spatial-MLLM (Wu et al., 2026a), SpaceR-7B (Ouyang et al., 2025), VST-7B (Yang et al., 2026a), VG-LLM-8B (Zheng et al., 2025), VLM-3R (Fan et al., 2026), SenseNova-SI-8B (Cai et al., 2026), and two SpatioLM (Wu et al., 2026c) variants. Several models share the same backbone model, enabling direction comparisons between base and spatially fine-tuned variants in Sec. 4.3: SpatialLadder and Spatial-MLLM are based on Qwen2.5-VL-3B; SpaceR, VST, and VG-LLM on Qwen2.5-VL-7B; VLM-3R on LLaVA-NeXT-Video-7B; SenseNova-SI on InternVL3-8B; and the two SpatioLM variants on InternVL3.5-8B and SenseNova-SI-8B, respectively.

Table B.1: **Evaluated models.** Checkpoint, backbone model, and parameter count of each evaluated model.

Category	Model	Checkpoint	Backbone model	# Params.
General-purpose	Qwen2.5-VL-3B	Qwen2.5-VL-3B-Instruct	–	3B
	Qwen2.5-VL-7B	Qwen2.5-VL-7B-Instruct	–	7B
	LLaVA-NeXT-Video-7B	LLaVA-NeXT-Video-7B-Qwen2	Qwen2-7B	7B
	InternVL3.5-8B	InternVL3_5-8B	Qwen3-8B	8B
	Qwen3-VL-8B	Qwen3-VL-8B-Instruct	–	8B
Spatially fine-tuned	SpatialLadder-3B	SpatialLadder-3B	Qwen2.5-VL-3B	3B
	Spatial-MLLM	Spatial-MLLM-v1.1-Instruct	Qwen2.5-VL-3B + VGGT	5B
	SpaceR-7B	SpaceR-SFT-7B	Qwen2.5-VL-7B	7B
	VST-7B	VST-7B-SFT	Qwen2.5-VL-7B	7B
	VG-LLM-8B	vgllm-qa-vggt-8b	Qwen2.5-VL-7B + VGGT	8B
	VLM-3R-7B	vlm-3r-llava-qwen2	LLaVA-NeXT-Video-7B + CUT3R	7B
	SenseNova-SI-8B	SenseNova-SI-1.1-InternVL3-8B	InternVL3-8B	8B
	SpatioLM-InternVL	SpatioLM-Understanding-InternVL3.5	InternVL3.5-8B + spatio-vision module	8B
	SpatioLM-SenseNova	SpatioLM-Understanding-SenseNovaSI	SenseNova-SI-8B + spatio-vision module	8B

Benchmarks and tasks. Tab. B.2 lists all 39 evaluated spatial tasks, along with their dataset sizes, answer formats, and evaluation metrics. *VSI-Bench* (Yang et al., 2025a) consists of egocentric indoor videos drawn from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021). We evaluate all eight tasks, totaling 5,130 items. *ReVSI* (Zhang et al., 2026b) re-annotates VSI-Bench-style questions across five indoor scan collections (*i.e.*, ARKitScenes, ScanNet, ScanNet++, 3RScan, MultiScan) using complete object annotations. We evaluate 13 question types, retaining counting, relative direction, relative distance, and room size tasks, including single/multiple objects, forward/backward $\times$ easy/hard, closest/farthest, and single/multiple rooms.

VSTI-Bench (Fan et al., 2026) is a spatio-temporal benchmark built on ScanNet videos. Its questions explicitly reference video frames, covering tasks such as camera displacement between frames, camera-object distance at a given frame, and object-relative position within a frame. *OSI-Bench* (Wu et al., 2026d) extends the scope of spatial reasoning evaluation from indoors to outdoors. It contains 997 outdoor, urban pedestrian-view videos with questions covering distances, displacements, speeds, ego-motion, 3D localization, and depth-aware counting. Videos are uniformly sampled to 32 frames for VSI-Bench, VSTI-Bench, and OSI-Bench. For ReVSI we use the 16-frame release for its ten original question types and the 32-frame release for the three additional types—room size, single and multiple, and route planning—which are not included in the 16-frame release. Across the four benchmarks, we evaluate 24,760 items in total: 5,130 from VSI-Bench, 5,128 from ReVSI, 5,736 from VSTI-Bench, and 8,766 from OSI-Bench, with each item scored by all 14 models.

Scoring. Numerical tasks are evaluated using mean relative accuracy (MRA) (Yang et al., 2025a), defined as the fraction of relative-error thresholds in $\{0.50, 0.55, \ldots, 0.95\}$ satisfied by a prediction, yielding a per-item score in $[0, 1]$. Multiple-choice tasks are scored by exact option match. All analyses use these per-item scores. For CST, we additionally average them across the 14 models and discretize the resulting values into ten equal-width bins to obtain the score variable Y.

Compute. All model inference for evaluation was conducted on a single node equipped with 8× NVIDIA RTX A6000 GPUs.

Table B.2: **Tasks.** Sizes, answer formats, and metrics of all 39 tasks. *Abbr.*: label used in Figs. 3 and 5; *Key*: task identifier in the released data; MRA: mean relative accuracy.

Benchmark	Task	Abbr.	Key	n	Answer	Metric
VSI-Bench	Appearance order	APP-ORD	obj_appearance_order	618	MCQ	accuracy
	Absolute distance	ABS-DIST	object_abs_distance	834	num.	MRA
	Object counting	CNT	object_counting	565	num.	MRA
	Relative direction	REL-DIR	object_rel_direction	968	MCQ	accuracy
	Relative distance	REL-DIST	object_rel_distance	710	MCQ	accuracy
	Object size	SIZE	object_size_estimation	953	num.	MRA
	Room size	ROOM	room_size_estimation	288	num.	MRA
	Route planning	ROUTE	route_planning	194	MCQ	accuracy
ReVSI	Absolute distance	ABS-DIST	object_abs_distance	909	num.	MRA
	Counting (multi)	CNT-M	object_counting_multiple	396	num.	MRA
	Counting (single)	CNT-S	object_counting_single	675	num.	MRA
	Rel. direction (bwd, easy)	RDIR-BE	object_rel_direction_backward_easy	156	MCQ	accuracy
	Rel. direction (bwd, hard)	RDIR-BH	object_rel_direction_backward_hard	198	MCQ	accuracy
	Rel. direction (fwd, easy)	RDIR-FE	object_rel_direction_forward_easy	255	MCQ	accuracy
	Rel. direction (fwd, hard)	RDIR-FH	object_rel_direction_forward_hard	237	MCQ	accuracy
	Rel. distance (closest)	RDIST-C	object_rel_distance_closest	443	MCQ	accuracy
	Rel. distance (farthest)	RDIST-F	object_rel_distance_farthest	289	MCQ	accuracy
	Object size	SIZE	object_size_estimation	1010	num.	MRA
	Room size (single)	ROOM-S	room_size_estimation_single	174	num.	MRA
	Room size (multi)	ROOM-M	room_size_estimation_multiple	128	num.	MRA
	Route planning	ROUTE	route_planning	258	MCQ	accuracy
VSTI-Bench	Camera displacement	CAM-DISP	camera_displacement	833	num.	MRA
	Camera movement dir.	CAM-DIR	camera_movement_direction	913	MCQ	accuracy
	Cam–obj abs. distance	CO-ABS	camera_obj_abs_dist	905	num.	MRA
	Cam–obj rel. dist. v1	CO-REL1	camera_obj_rel_dist_v1	91	MCQ	accuracy
	Cam–obj rel. dist. v2	CO-REL2	camera_obj_rel_dist_v2	493	MCQ	accuracy
	Cam–obj rel. dist. v3	CO-REL3	camera_obj_rel_dist_v3	856	MCQ	accuracy
	Obj–obj rel. pos. (L/R)	OO-LR	obj_obj_relative_pos_lr	605	MCQ	accuracy
	Obj–obj rel. pos. (N/F)	OO-NF	obj_obj_relative_pos_nf	556	MCQ	accuracy
	Obj–obj rel. pos. (U/D)	OO-UD	obj_obj_relative_pos_ud	484	MCQ	accuracy
OSI-Bench	Absolute displacement	ABS-DISP	absolute_displacement	1058	num.	MRA
	Absolute distance	ABS-DIST	absolute_distance	997	num.	MRA
	Absolute speed	SPEED	absolute_speed	1058	num.	MRA
	Depth-aware counting	CNT-D	depth_aware_counting	998	num.	MRA
	3D localization	LOC-3D	object_3d_localization	994	num.	MRA
	Relative direction	REL-DIR	relative_direction_categorical_ordinal	673	MCQ	accuracy
	Relative distance	REL-DIST	relative_distance	996	MCQ	accuracy
	Qualitative ego-motion	QUAL-EM	trajectory_description	996	MCQ	accuracy
	Quantitative ego-motion	QUAN-EM	trajectory_length	996	num.	MRA

C Additional Related Work

This section extends Sec. 2 with additional literature relevant to the methods and evaluation setting considered in this work.

C.1 Spatial Reasoning in Vision-Language Models

There is rapidly growing interest in improving the spatial reasoning capabilities of VLMs. Building on the data-centric approach introduced by SpatialVLM (Chen et al., 2024), subsequent work extends this direction with richer spatial supervision, broader task coverage, and scalable training pipelines to build more spatially capable backbones (Yang et al., 2026a; Zhang et al., 2025; Daxberger et al., 2025; Chen et al., 2026b; Li et al., 2026; AI et al., 2025; Yin et al., 2026; Lian et al., 2026; Liu et al., 2026; Brown et al., 2025a; Deng et al., 2026). On the architectural side, 3DThinker (Chen et al., 2026c), Spatial-MLLM (Wu et al., 2026a), related approaches introduce architectures that integrate geometric cues more directly into image- and video-based VLMs (Zhao et al., 2025; Zhang et al., 2026a; Fan et al., 2026; Zheng et al., 2025; Yang et al., 2026b). Another line of work adopts tool-based or agentic approaches, combining the high-level planning and reasoning capabilities of VLMs with grounded visual or geometric tools (Ma et al., 2024; Lee et al., 2025b; Cho et al., 2026; Chen et al., 2026a; Zhou et al., 2026; Luo et al., 2026).

Orthogonally, there is a dense literature on developing benchmarks and datasets for evaluating spatial reasoning in VLMs. Notable efforts include RoboSpatial (Song et al., 2025), CV-Bench (Tong et al., 2024), and SPAR-Bench (Zhang et al., 2025). These benchmarks cover diverse spatial capabilities, including metric estimation, relation reasoning, embodied grounding, navigation, and long-horizon video understanding Wu et al. (2026b); Lin et al. (2026); Yang et al. (2026c; 2025b); Yu et al. (2025); Lin et al. (2025).

C.2 Statistical Disentanglement and Reweighting

Our framework builds on statistical methods for analyzing correlated factors in observational data. Information-theoretic measures provide a natural foundation: mutual information (Cover et al., 1991; Shannon, 1948) quantifies dependence between a factor and an outcome, while conditional mutual information measures the dependence that remains after conditioning on another factor. Related ideas appear in causal and statistical adjustment methods, including mediation analysis (VanderWeele, 2015), propensity-score and inverse-probability weighting (Rosenbaum & Rubin, 1983; Hernán & Robins, 2010), and the g-formula (Robins, 1986). In multimodal research, such tools have been used to analyze language priors in VQA (Niu et al., 2021; Cadene et al., 2019; Clark et al., 2019) and, more recently, to diagnose spurious correlations and reasoning failures in VLMs (Xu et al., 2026; Komanduri et al., 2025; Yang et al., 2025c). Our work applies these ideas to analyze spatial reasoning benchmark composition. CST uses marginal and conditional mutual information to identify task-agnostic factors whose associations with scores persist after adjustment for entangled factors. FBS then uses stable balancing weights (Zubizarreta, 2015) to balance the diagnosed factor levels while controlling weight dispersion. Together, they provide a framework for diagnosing and mitigating composition sensitivity in spatial reasoning evaluation.

D Details on Task-Agnostic Factors

Catalog. Every benchmark item is annotated with metadata that can be measured directly from the item, grouped into seven factor families corresponding to those used in Fig. 3b:

- **Ground-truth answer value**: the numerical answer (*e.g.*, count, distance in m, size in cm, area in m^2, or speed in m/s) or, for multiple-choice tasks, the released answer label. For VSTI-Bench, this label corresponds to the answer *class* (*e.g.*, left/right, near/far, camera-motion direction, or object name), whereas for VSI-Bench, ReVSI, and OSI-Bench it is the option letter, capturing potential option-position effects. Redundant copies of the same target (*e.g.*, numeric value, answer letter, and answer index) are removed by de-duplication as described below.
- **Object type**: a difficulty proxy for the queried object category, defined using the per-item score of a fixed reference model (Qwen2.5-VL-7B) and discretized with the same tail-aware rule as other numeric factors. For tasks with MRA scores, the proxy has four bins; for all-or-nothing scoring, it is binary and may collapse to a single level, in which case it is excluded from that task's ranking. This operationalization follows our earlier study and is convenient because it is available on every benchmark, but it is not independent of the evaluated pool (Sec. 3.2); a category-level prior is a natural replacement.
- **Scene / room**: metadata describing the scene context, including scan source (VSI-Bench); room type from dataset labels where available, otherwise estimated with a CLIP-based room classifier (Radford et al., 2021), together with an additional VLM-based estimate (VSI-Bench; VLM

estimate only for ReVSI); VLM-rated scene complexity (VSI-Bench, ReVSI); scene area type and number of annotated objects (ReVSI, VSTI-Bench); number of distinct object categories (ReVSI, VSTI-Bench, OSI-Bench); and VLM-predicted scene type and queried depth threshold (OSI-Bench).

- **Question / options**: question length in words; number of options; number of queried objects.
- **Temporal / frames**: duration (VSI-Bench, OSI-Bench), frame count, scene-cut count, dominant optical-flow direction and CLIP-embedding diversity of frames (VSI-Bench), inter-frame difference (VSI-Bench, ReVSI); referenced frame index and frame gap, absolute and relative (VSTI-Bench); queried time and time gap, absolute and relative (OSI-Bench).
- **Appearance**: mean brightness, contrast and sharpness (all benchmarks), edge density (VSI-Bench, ReVSI) and brightness dynamic range across the sampled frames (VSI-Bench).
- **File / resolution**: file size, resolution, width and height, aspect ratio, frame rate.

Across benchmarks, 37-38 metadata keys are available (38 for VSI-Bench and VSTI-Bench; 37 for ReVSI and OSI-Bench). After screening, 11-31 factors remain per task, reduced to 8-28 after de-duplication.

Screening. A metadata key is a candidate factor for a task if it has at least 50 non-missing values, is not an identifier, list, or constant, and has at most 40 categorical levels; bookkeeping fields that only record which annotator model produced another field are excluded as well. Keys whose values are at least 90% numeric are treated as numeric and binned; the rest are categorical and used natively.

Binning. Numeric factors are discretized with a tail-aware four-bin rule: the top 20% of the values (above the 80th percentile) form one bin and the remaining range is split into three equal-width bins. This rule keeps the discriminative upper tail of long-tailed factors (answer values, durations, file sizes) as a separate bin while avoiding degenerate bins on heavily tied factors such as small integer counts; when fewer than four distinct edges result, the number of bins is reduced accordingly. Categorical factors are used at their native level set. The score S is the per-item score averaged over models, discretized into ten equal-width bins on $[0, 1]$.

De-duplication. Benchmark metadata contain copies and recodings of the same variable (*e.g.*, a numeric copy of the ground truth, the answer letter and its index, frame count and sampling stride, a raw and a cleaned room-type field). Such duplicates would occupy top-10 slots, make each other look like surface correlates (*i.e.*, $I(Y; X_f \mid X_g) = 0$ when X_g is a copy of X_f), and constrain the same axis twice during balancing. We remove (i) a fixed list of known copies (numeric ground-truth copy, answer letter and index, random-baseline accuracy, sampling stride, duplicate duration field, and raw-output fields) and (ii) any factor whose normalized mutual information with an already retained factor is at least 0.95, keeping the one with the higher marginal MI with the score.

E Details on Composition Sensitivity Test (CST)

In this section, we provide additional details on the design choices behind our Composition Sensitivity Test (CST) in Sec. 3.2, along with further analysis results.

Choice of c. The direct contributor set depends on both the magnitude threshold c and the persistence tolerance r. We fix $r = 1.5$ and calibrate c using two target-free criteria: permutation-null calibration and bootstrap stability (Tab. E.1).

Permutation null. For each task, we shuffle score bins across items 200 times and compute the 95th percentile of each top-10 factor's corrected MI under this null. We require γ_t to exceed this percentile for every selected direct contributor. Increasing c from 0.015 to 0.02 reduces the number of tasks violating this criterion from ten to three. The remaining exceptions are small tasks: ReVSI single-room size ($N = 174$), ReVSI multiple-room size ($N = 128$), and VSTI-Bench camera-object relative distance v1 ($N = 91$). Their diagnoses therefore require additional caution.

Bootstrap stability and coverage. Across 100 item-level bootstrap resamples, the mean Jaccard similarity (Jaccard, 1901) between the resampled and full-data direct sets increases from 0.44 at $c = 0.005$ to 0.66 at $c = 0.02$. Higher thresholds yield smaller stability gains (0.69–0.73), while the number of balanced tasks falls from 29 at $c = 0.02$ to 22 at $c = 0.05$. We select $c = 0.02$: the smallest tested value satisfying the null criterion for every task with $N \geq 200$, while retaining 29 of 39 tasks for balanced evaluation.

Table E.1: **Threshold sweep** ($c \in [0.005, 0.05]$, all four benchmarks, 14 models, 39 tasks). Bootstrap statistics use 100 item resamples; the permutation null uses 200 score shuffles per factor. "Below its null" counts tasks with at least one direct contributor whose γ_t is below the 95th percentile of its chance MI.

c	0.005	0.0075	0.01	0.0125	0.015	**0.02**	0.025	0.03	0.04	0.05
Balanced tasks (of 39)	37	37	35	34	33	29	29	28	26	22
VSI-Bench	8	8	8	8	8	6	6	6	5	3
ReVSI	13	13	12	11	11	9	9	8	7	6
VSTI-Bench	7	7	7	7	6	6	6	6	6	5
OSI-Bench	9	9	8	8	8	8	8	8	8	8
Mean # direct contributors / task	3.62	3.18	2.82	2.33	2.15	1.85	1.59	1.33	1.00	0.87
Bootstrap Jaccard vs. full data	0.44	0.48	0.53	0.56	0.60	0.66	0.69	0.71	0.72	0.73
Pairwise bootstrap Jaccard	0.63	0.63	0.64	0.65	0.66	0.68	0.69	0.71	0.74	0.75
Top-10 factors passing γ by chance (%)	50.3	37.7	30.5	24.9	19.5	9.6	7.0	4.3	1.3	0.0
Tasks with a direct factor below its null	23	17	12	11	10	3	3	2	1	0

Finite-sample estimation. For each task, the score variable Y is the per-item score averaged over the 14 models and discretized into ten equal-width bins on $[0, 1]$. After de-duplication in App. D, we retain the ten factors with the highest marginal MI as the candidate set. We estimate MI and CMI from empirical contingency tables using the Miller-Madow correction (Miller, 1955),

$$\widehat{I}_{\mathrm{MM}}(Y; X_f) = \max\left(0,\ \widehat{I}(Y; X_f) - \frac{(\widehat{K}_S - 1)(\widehat{K}_f - 1)}{2N}\right),$$

where $\widehat{K}_S$ and $\widehat{K}_f$ are the numbers of observed levels. For conditional MI (CMI), we apply the same correction within each X_g stratum and average the corrected estimates by stratum size. No stratum is dropped; sparse strata are handled through the bias correction. The residual dependence of factor f is the mean corrected CMI over the remaining task-agnostic factors, while skipping pairs with fewer than 50 jointly observed items or a constant adjustment factor.

Robustness to joint conditioning. In Eq. (4), CST calculate residual dependence by conditioning on one factor at a time and averaging across adjustment factors. To test whether higher-order entanglement changes the diagnosis, we repeat the analysis while conditioning jointly on $m \in \{2, 3\}$ factors and average over all combinations of adjustment factors (Tab. E.2). Of the 72 direct factor-task pairs identified at $c = 0.02$, 60 (83%) remain direct under two-factor conditioning, including 38 of 43 answer-value and object-type contributors (88%). Importantly, none falls below the magnitude threshold γ_t: the 12 changes are instead caused by the residual-to-marginal dependence ratio moving outside the persistence band, with nine reclassified as *suppressed* and three as *surface correlates*.

Table E.2: **Higher-order conditioning** ($c = 0.02$).

m	1	2	3
Retained (%)	100	83	51

Under three-factor conditioning, 37 of 72 pairs (51%) remain direct, and seven fall below γ_t. At the same time, the mean number of occupied conditioning strata increases from 4.4 at $m = 1$ to 15.6 at $m = 2$ and 44.1 at $m = 3$, leaving substantially fewer samples per stratum, particularly for smaller tasks. We therefore use single-factor adjustment for the main diagnosis and treat higher-order conditioning as a sensitivity analysis in CST: the diagnosis is largely preserved at $m = 2$, while estimates become increasingly data-limited at $m = 3$. Larger benchmarks in the future should enable more reliable multi-factor conditioning and, in turn, more robust CST diagnosis.

Full per-task results. Figs. G.6 to G.9 show the full per-task CST results for all four benchmarks, visualizing the top-10 factors and their corresponding sensitivity regimes.

F Details on Factor-Balanced Score (FBS)

In this section, we provide additional details on the design choices behind our factor balancing in FBS in Sec. 4.1.

Cells, targets, and feasibility. For a task with direct contributors $\mathcal{G}$, we group items by the joint configurations of their (binned) factor levels. With K_f levels for factor f, there are at most $\prod_f K_f$ cells, of which 2–143 are occupied across our tasks (Tab. G.2). Our target assigns a uniform marginal

distribution over the occupied levels of each diagnosed factor. When multiple factors are balanced jointly (*i.e.*, $|\mathcal{G}| \geq 2$), these uniform marginals may be infeasible because some joint cells are unobserved. In such cases, we project the uniform target onto the closest feasible marginals in squared error using a small quadratic program over cell weights, and use the resulting target for all reweighting methods. Uniform marginals are exactly feasible on 27 of the 29 balanced tasks. The two exceptions are VSI-Bench's appearance order and ReVSI's room size (multi), with mean total-variation distances of 0.09 and 0.02 from the uniform target, respectively.

Stable balancing weights (SBW). As described in Sec. 4.1, we use stable balancing weights (SBW) (Zubizarreta, 2015) for reweighting. Given a target marginal weight $q_{f,v}$ for each level v of diagnosed factor f, SBW chooses weights by solving

$$\min_{\{w_i \geq 0\}} \sum_{i \in \mathcal{I}} w_i^2 \quad \text{s.t.} \quad \sum_{i \in \mathcal{I}} w_i = 1, \; \left| \sum_{i \in \mathcal{I}} w_i \, \mathbf{1}[X_{i,f} = v] - q_{f,v} \right| \leq \delta, \; \forall f \in \mathcal{G}_t, \; v \in \mathcal{V}_f. \tag{6}$$

We use $\delta = 0.05$ for all benchmarks. Because the weights sum to one, minimizing $\sum_i w_i^2$ equivalently maximizes the effective sample size, $\text{ESS} = 1/\sum_i w_i^2$, while allowing small deviations from the balancing targets. At least one balance constraint is active on 27 of the 29 balanced tasks, indicating that the tolerance is typically binding. Across tasks, the resulting ESS ranges from 16% to 63% of the original sample size (34% on average). The maximum relative item weight, $\max_i |\mathcal{I}| w_i$, ranges from 5 to 64 (28 on average). We compare SBW with alternative reweighting schemes below.

Choice of δ. The balanced score depends on the tolerance δ in the SBW formulation of Eq. (6), which controls how much exact balance is traded for stable weights. We fix $c = 0.02$ and calibrate δ using two criteria: stability under composition resampling and fidelity to exact balance (Tab. F.1).

Stability. Under the composition-resampling protocol of the reweighting ablation (Tab. F.3), stability is U-shaped in δ: exact balance ($\delta = 0$) forces large weights onto sparse levels, whereas a loose tolerance lets the weights drift back toward the observed composition. Stability improves up to $\delta = 0.05$ (rank stability $0.558 \to 0.594$; score standard deviation $0.0668 \to 0.0585$) and plateaus over $[0.05, 0.10]$.

Fidelity. The cost of the tolerance grows with δ. The balanced scores depart from those under exact balance by 0.023 at $\delta = 0.05$ but by 0.052 at $\delta = 0.3$, where the effective sample size has risen from 23% to 86% of the items (*i.e.*, the weights have nearly returned to the observed composition.) We select $\delta = 0.05$: the smallest tested value that reaches the stability plateau while keeping the balanced scores close to exact balance.

Table F.1: **Tolerance sweep** ($\delta \in [0, 0.3]$, $c = 0.02$, 14 models, 29 balanced tasks). Stability statistics use the $B = 200$ composition resamples of Tab. F.3; ESS is the effective sample size; MAE vs. δ=0 is the mean absolute difference of the balanced scores from those under exact marginal balance.

δ	0	0.01	0.02	0.03	**0.05**	0.075	0.1	0.15	0.2	0.3
Rank stability $\bar{\tau} \uparrow$	0.558	0.572	0.582	0.588	0.594	0.596	0.594	0.586	0.572	0.544
Score s.d. $\sigma_s \downarrow$	0.0668	0.0633	0.0612	0.0598	0.0585	0.0586	0.0599	0.0645	0.0701	0.0800
ESS (% of items)	23.0	25.1	27.3	29.6	34.5	40.7	46.9	59.2	71.4	86.0
MAE vs. δ=0	0.0000	0.0080	0.0127	0.0163	0.0233	0.0297	0.0348	0.0415	0.0466	0.0515

Ablation on reweighting schemes. Here, we compare multiple options for reweighting. Each scheme assigns item weights w_i used to compute the factor-balanced score $S_m^{\text{bal}} = \sum_{i \in \mathcal{I}} w_i s_{m,i}$ in Eq. (5), but differs in the balance constraints it imposes and how it selects among feasible weights (Tab. F.2). Note that the observed aggregate score S_m^{obs} corresponds to uniform weighting over all benchmark items and therefore follows the benchmark's empirical composition. *Joint-uniform post-stratification* ("macro") assigns equal total weight to each occupied joint factor cell; under ragged support, however, this does not generally yield uniform factor marginals. *Inverse-probability weighting (IPW)* assigns weights proportional to $\prod_f 1/\hat{p}_f(c_{i,f})$, providing a one-shot approximation that is exact only when the factors are independent. *Raking* (Deville & Särndal, 1992) exactly matches the target marginals while minimizing KL divergence, while *entropy balancing* (Hainmueller, 2012) solves an equivalent optimization in dual form. For categorical marginal constraints, the two yield equivalent weights, as reflected by our results to three decimal places. *Multilevel calibration* (Ben-Michael et al., 2024) also exactly matches the target marginals but selects minimum-variance weights.

Stable balancing weights (SBW) (Zubizarreta, 2015), which we use in our main experiments, instead allow a tolerance δ on the balance constraints while minimizing weight variance. For tasks with a single CST-diagnosed factor, the marginal and joint targets coincide, so all exact-balancing schemes yield identical weights. Differences arise only for the 22 tasks with two or more diagnosed factors.

Table F.2: **Reweighting schemes.** All target uniform marginals (or the uniform joint) of the CST direct contributors.

Scheme	Constraint (target)	Objective	Key property
Raw (observed)	none ($W_c \propto n_c$)	–	current aggregate
Joint-uniform (macro)	equal weight per occupied cell	– (unique)	marginals not uniform on ragged support
IPW	marginal uniform, one shot $\prod_f 1/\hat{p}_f$	– (unique)	assumes independent factors
Raking / IPF	marginal uniform (exact)	min KL	standard survey calibration
Entropy balancing	marginal uniform (exact)	min KL (dual)	equals raking for categorical marginals
SBW (δ=0.05)	marginal within $\pm\delta$	min variance (= max ESS)	FBS estimator
Multilevel calibration	marginal uniform (exact)	min variance	exact solution with smallest weights

Ablation results. Tab. F.3 reports four stability criteria, effective sample size (ESS), maximum relative weight, and maximum deviation from the target marginals for all seven schemes, both by benchmark and pooled across the 29 balanced tasks. All methods are evaluated on the same $B = 200$ composition resamples, enabling paired comparisons. Considering stability alone, SBW performs best on 8 of the 16 benchmark-criterion combinations, while joint-uniform post-stratification performs best on the remaining 8, including all four criteria on VSI-Bench, three on ReVSI, and the flip rate on VSTI-Bench. However, joint-uniform post-stratification does not in general achieve the intended balance target: on multi-factor tasks, its marginal deviation reaches 0.15 (marked †), because equal weighting over occupied joint cells does not imply uniform factor marginals under ragged support. Its apparent stability therefore reflects, in part, an estimand whose weights change little under composition resampling.

Restricting the comparison to schemes that satisfy the target marginals within δ on every task—raking, entropy balancing, multilevel calibration, and SBW—SBW achieves the highest stability on all 16 benchmark-criterion combinations. It also retains the largest ESS on every benchmark (24–40%, compared with 15–28% for the other feasible schemes) and yields the smallest maximum relative weight. In contrast, IPW is the only reweighting scheme that is less stable than the observed score when pooled across tasks, with rank stability of 0.48 versus 0.50. These conclusions are unchanged at $c = 0.015$ and $c = 0.025$, where SBW again achieves the highest stability on all 16 feasible benchmark-criterion comparisons. Based on this ablation, we use SBW to compute FBS throughout the main experiments.

G Full Experiment Results

Composition robustness experiment protocol. We provide the full protocol for the robustness experiment in Sec. 4 (Fig. 6). For each of the 29 tasks with at least one direct contributor, we perturb the composition along the *top-1 direct contributor*, defined as the direct contributor with the highest marginal MI. The selected factor is the answer value on 20 tasks, the object type on 6 tasks, and brightness dynamic range, question length, and mean brightness on one task each. Items are ranked by the selected factor's value, using a fixed category ordering for categorical factors. Each item receives a sampling probability proportional to a Beta density evaluated at its normalized rank $u_i \in (0, 1)$, with mean $0.5 + 0.35\lambda$ and concentration 2. At $\lambda = 0$, the density is uniform, corresponding to uniform sampling over items and preserving the benchmark composition in expectation. Negative and positive values of λ emphasize lower- and higher-ranked items, respectively. We sweep λ over 200 equally spaced values in $[-1, 1]$, each defining a composition setting. For each setting, we draw 20 samples of n' items with replacement from the same task, keeping the underlying item pool fixed. Here, n denotes the total number of items in the task. The main-text experiment uses $n' = n/2$, while Tab. G.1 additionally reports results for $n' = n/3$ and $n' = n/4$.

For each sample, we compute S_m^{obs} as the unweighted mean of the per-item scores $s_{m,i}$ and S_m^{bal} using Eq. (5). To compute the latter, we retain the full task's direct-contributor set and bin edges but re-estimate the SBW item weights w_i on the sample with $\delta = 0.05$. These weights depend only on the sample's composition and are shared across all evaluated models. For each composition setting, we average the 20 replicate score vectors separately for the observed and balanced scores, then rank models using these composition-wise mean scores.

Table F.3: **Reweighting-scheme ablation under composition resampling** ($c = 0.02$, 14 models, $B = 200$). Bold: best balancer per column and benchmark; underline: second best (Raw excluded from both). $\max|\text{dev}|$: largest deviation of a factor marginal from its target; † marks schemes that miss the target by more than δ on at least one task of the benchmark.

Benchmark	Scheme	$\bar{\tau}\uparrow$	Top-1↑	$\sigma_s\downarrow$	Flip↓	ESS%↑	$\max w\downarrow$	$\max\|\text{dev}\|$
VSI-Bench (6)	Raw (observed)	0.647	0.576	0.0826	0.251	100.0	1.0	0.364
	Joint-uniform (macro)	**0.673**	**0.553**	**0.0507**	**0.237**	18.8	33.0	0.144†
	IPW	0.499	0.443	0.1255	0.363	17.2	73.7	0.310†
	Raking / IPF	0.611	0.526	0.0633	0.284	19.3	41.3	0.001
	Entropy balancing	0.610	0.525	0.0636	0.286	19.2	41.4	0.000
	SBW (δ=0.05)	0.668	0.552	0.0515	0.244	**34.8**	**23.5**	0.050
	Multilevel calibration	0.611	0.526	0.0632	0.286	21.0	36.4	0.000
ReVSI (9)	Raw (observed)	0.596	0.496	0.0853	0.305	100.0	1.0	0.353
	Joint-uniform (macro)	**0.645**	**0.559**	0.0595	**0.261**	29.1	27.3	0.118†
	IPW	0.513	0.436	0.1082	0.356	25.5	41.0	0.225†
	Raking / IPF	0.589	0.478	0.0694	0.300	26.7	32.1	0.001
	Entropy balancing	0.587	0.477	0.0697	0.302	26.6	32.3	0.000
	SBW (δ=0.05)	0.624	0.537	**0.0594**	0.275	**39.6**	**19.7**	0.049
	Multilevel calibration	0.586	0.480	0.0694	0.302	27.5	28.4	0.000
VSTI-Bench (6)	Raw (observed)	0.340	0.309	0.1181	0.503	100.0	1.0	0.372
	Joint-uniform (macro)	0.486	0.405	0.0766	**0.370**	25.5	34.2	0.070†
	IPW	0.398	0.330	0.1151	0.448	23.8	38.8	0.141†
	Raking / IPF	0.458	0.423	0.0823	0.397	25.1	34.3	0.000
	Entropy balancing	0.458	0.420	0.0826	0.398	25.0	34.4	0.000
	SBW (δ=0.05)	**0.486**	**0.433**	**0.0734**	0.372	**39.3**	**21.2**	0.049
	Multilevel calibration	0.459	0.422	0.0825	0.396	25.9	31.3	0.000
OSI-Bench (8)	Raw (observed)	0.402	0.506	0.1123	0.460	100.0	1.0	0.423
	Joint-uniform (macro)	0.578	0.703	0.0519	0.314	12.2	66.5	0.095†
	IPW	0.492	0.589	0.0828	0.374	14.6	88.9	0.222†
	Raking / IPF	0.575	0.701	0.0522	0.313	15.5	70.9	0.000
	Entropy balancing	0.574	0.700	0.0522	0.314	15.5	71.0	0.000
	SBW (δ=0.05)	**0.598**	**0.719**	**0.0497**	**0.298**	**24.1**	**44.4**	0.050
	Multilevel calibration	0.577	0.713	0.0519	0.313	16.9	59.1	0.000
All (29 tasks)	Raw (observed)	0.500	0.477	0.0990	0.378	100.0	1.0	0.379
	Joint-uniform (macro)	**0.599**	0.566	0.0591	**0.293**	21.6	40.7	0.107†
	IPW	0.481	0.458	0.1062	0.382	20.4	60.5	0.224†
	Raking / IPF	0.562	0.538	0.0661	0.321	21.7	45.2	0.001
	Entropy balancing	0.561	0.537	0.0663	0.322	21.7	45.3	0.000
	SBW (δ=0.05)	0.597	**0.569**	**0.0580**	0.295	**34.3**	**27.6**	0.049
	Multilevel calibration	0.562	0.542	0.0660	0.321	22.9	39.1	0.000

We define *ranking stability* as the mean Kendall's τ over 3,000 randomly sampled pairs of composition-wise leaderboards. We define *score stability* as $1 - \text{SD}/\text{SD}_{\max}$, where SD is the standard deviation of each model's composition-wise mean scores across the 200 settings, averaged over models. The fixed normalization constant $\text{SD}_{\max} = 0.200$ is the largest model-wise standard deviation observed across all models, tasks, sample sizes, and scoring methods. Both metrics are computed separately for S_m^{obs} and S_m^{bal}, first per task and then averaged over the evaluated tasks within each benchmark. Tasks without a direct contributor are excluded because FBS coincides with the observed score on these tasks. Averaging over 20 replicates reduces within-composition sampling noise. The estimated standard error of a composition-wise mean, computed as the within-composition score standard deviation divided by $\sqrt{20}$, is 0.006–0.007 on average. For comparison, the between-composition score standard deviation averages 0.061 for the observed score and 0.020 for FBS.

Tab. G.1 reports both stability metrics for all three sample sizes. Balancing improves benchmark-averaged ranking stability on every benchmark at every tested sample size, with comparable gains from $n/2$ to $n/4$. Thus, the robustness gains from FBS persist across the tested sample sizes.

Per-task overview. Tab. G.2 summarizes the results for each task, including the CST-diagnosed direct contributors, the number of occupied joint factor cells, ranking and score stability under composition resampling before and after balancing, the number of models whose rank changes, whether the top-ranked model changes, and the mean absolute composition advantage $|\Delta S_m|$. Two patterns emerge. First, tasks with no direct contributor consistently belong to the relative direction, relative distance, and relative position families across the four benchmarks. Second, the largest mean absolute composition advantages occur primarily in counting tasks—VSI-Bench's counting, ReVSI's

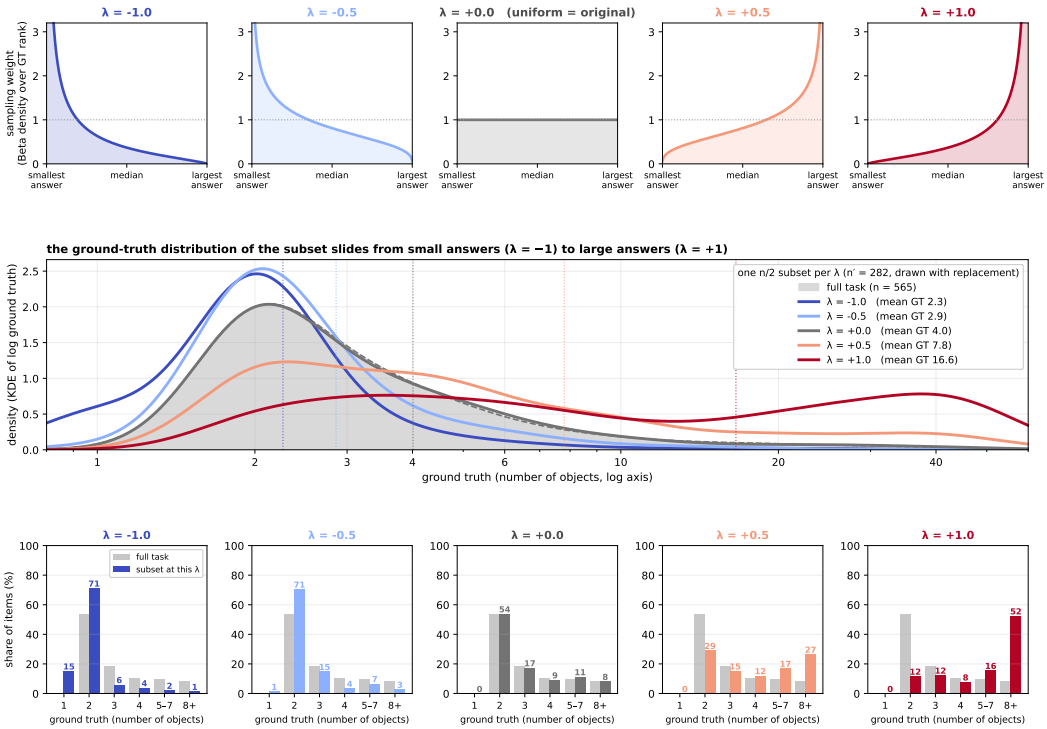

Figure G.1: **Composition resampling along the top-1 direct contributor.** Example: VSI-Bench's object counting, with answer value as the selected factor. **Top:** Sampling weights versus normalized item rank along the factor for five values of λ, using a Beta density with mean $0.5 + 0.35\lambda$ and concentration 2. At $\lambda = 0$, uniform item sampling preserves the original composition in expectation. **Middle:** Ground-truth answer distributions for one $n/2$-item sample drawn with replacement at each λ, compared with the full task (kernel density estimates displayed on a logarithmic axis). **Bottom:** Proportions of items at each ground-truth count in the same samples. Increasing λ shifts the composition from smaller to larger answer values while keeping the underlying item pool fixed.

Table G.1: **Stability under composition resampling by subset size.** Mean over the balanced tasks of each benchmark; 200 compositions along the top-1 direct contributor $\times$ 20 subsets of n' items. Ranking stability: mean Kendall's τ over 3,000 composition pairs. Score stability: $1 - \mathrm{SD}/\mathrm{SD}_{\max}$ with $\mathrm{SD}_{\max} = 0.200$. Bold: higher of Raw and FBS.

Subset	Benchmark	Ranking stability $\bar{\tau}$ ↑		Score stability ↑	
		Raw	FBS	Raw	FBS
$n/2$	VSI-Bench	0.78	**0.85**	0.70	**0.89**
	ReVSI	0.71	**0.85**	0.71	**0.91**
	VSTI-Bench	0.52	**0.78**	0.66	**0.89**
	OSI-Bench	0.55	**0.85**	0.69	**0.91**
$n/3$	VSI-Bench	0.77	**0.84**	0.70	**0.88**
	ReVSI	0.71	**0.84**	0.70	**0.90**
	VSTI-Bench	0.52	**0.76**	0.66	**0.88**
	OSI-Bench	0.55	**0.84**	0.69	**0.91**
$n/4$	VSI-Bench	0.77	**0.83**	0.70	**0.88**
	ReVSI	0.71	**0.83**	0.70	**0.89**
	VSTI-Bench	0.51	**0.75**	0.66	**0.88**
	OSI-Bench	0.55	**0.84**	0.69	**0.90**

multi-object counting, and OSI-Bench's depth-aware counting—as well as OSI-Bench's qualitative ego-motion, whose four-way answer distribution is strongly skewed.

Re-ranking charts for all balanced tasks. Figs. G.2 to G.5 show the observed-to-balanced re-ranking of all 14 models on every balanced task.

Table G.2: **Per-task results** ($c = 0.02$, SBW $\delta = 0.05$, $B = 200$). $\bar{\tau}$ and σ_s: rank stability and score standard deviation under composition resampling, observed → balanced. *Ranks ch.*: models (of 14) whose rank changes. $\overline{|\Delta S_m|}$: mean absolute composition advantage over models.

Bench.	Task	n	Format	Direct contributors	Cells	$\bar{\tau}$ obs→bal	σ_s obs→bal	Ranks ch.	Top-1 flip	$\overline{\lvert\Delta S_m\rvert}$
VSI-Bench	Appearance order	618	MCQ	brightness range, room type, room type (VLM), +4 more	32	0.88→0.95	0.033→0.012	6	–	0.022
	Abs. distance	834	num.	GT value, object type	15	0.81→0.92	0.053→0.016	4	–	0.030
	Object counting	565	num.	GT value, object type	10	0.65→0.65	0.107→0.046	13	✓	0.127
	Rel. direction	968	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Rel. distance	710	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Object size	953	num.	object type, GT value, category freq.	48	0.86→0.94	0.073→0.024	2	–	0.009
	Room size	288	num.	GT value, object type, room type, file size	143	0.85→0.91	0.050→0.019	8	✓	0.031
	Route planning	194	MCQ	question length, GT value	15	0.60→0.71	0.045→0.019	6	–	0.011
ReVSI	Abs. distance	909	num.	object type, GT value	16	0.88→0.98	0.036→0.008	2	–	0.039
	Counting (multi)	396	num.	object type, GT value, #scene objects	37	0.83→0.82	0.074→0.022	12	–	0.113
	Counting (single)	675	num.	GT value, object type	11	0.39→0.82	0.081→0.018	13	✓	0.029
	Rel. dir. (bwd, easy)	156	MCQ	GT value	3	0.55→0.67	0.039→0.017	0	–	0.000
	Rel. dir. (bwd, hard)	198	MCQ	object type	2	0.83→0.92	0.027→0.012	11	✓	0.036
	Rel. dir. (fwd, easy)	255†	MCQ	GT value, scene area type	6	0.64→0.72	0.034→0.014	2	–	0.001
	Rel. dir. (fwd, hard)	237	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Rel. dist. (closest)	443	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Rel. dist. (farthest)	289	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Object size	1010	num.	object type, GT value, question length	32	0.83→0.93	0.054→0.015	2	–	0.011
	Room size (multi)	128	num.	GT value, object type, #scene objects, +5 more	122	0.82→0.95	0.072→0.022	6	–	0.049
	Room size (single)	174	num.	GT value, object type, room type (VLM)	56	0.64→0.83	0.110→0.027	9	✓	0.015
	Route planning	258	MCQ	∅ (not balanced)	–	–	–	–	–	–
VSTI-Bench	Camera displacement	833	num.	GT value	4	0.45→0.89	0.047→0.011	2	–	0.019
	Camera move dir.	913†	MCQ	GT value, #options, #scene objects	31	0.29→0.63	0.084→0.018	9	–	0.040
	Cam-obj abs. dist.	905	num.	GT value, object type	16	0.43→0.66	0.078→0.025	10	–	0.021
	Cam-obj rel. dist. v1	91	MCQ	brightness, sharpness, file size, contrast	29	0.81→0.90	0.042→0.028	11	✓	0.037
	Cam-obj rel. dist. v2	493	MCQ	GT value	22	0.62→0.64	0.048→0.025	14	✓	0.079
	Cam-obj rel. dist. v3	856	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Obj-obj rel. pos. (l/r)	605	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Obj-obj rel. pos. (n/f)	556	MCQ	GT value	2	0.54→0.94	0.104→0.025	0	–	0.000
	Obj-obj rel. pos. (u/d)	484	MCQ	∅ (not balanced)	–	–	–	–	–	–
OSI-Bench	Abs. displacement	1058	num.	GT value, object type	12	0.50→0.85	0.035→0.015	11	–	0.026
	Abs. distance	997	num.	GT value, object type	16	0.66→0.94	0.034→0.017	7	✓	0.021
	Abs. speed	1058	num.	GT value, object type	16	0.44→0.85	0.047→0.011	11	–	0.033
	Depth-aware counting	998	num.	object type, GT value	11	0.70→0.95	0.065→0.006	9	✓	0.080
	3D localization	994	num.	GT value, object type	16	0.52→0.91	0.059→0.024	11	–	0.026
	Rel. direction	673	MCQ	GT value	4	0.25→0.52	0.047→0.014	8	–	0.021
	Rel. distance	996	MCQ	∅ (not balanced)	–	–	–	–	–	–
	Qualitative ego-motion	996	MCQ	GT value	4	0.64→0.87	0.157→0.039	9	✓	0.098
	Quantitative ego-motion	996	num.	GT value, duration, file size	49	0.72→0.93	0.043→0.016	9	–	0.028

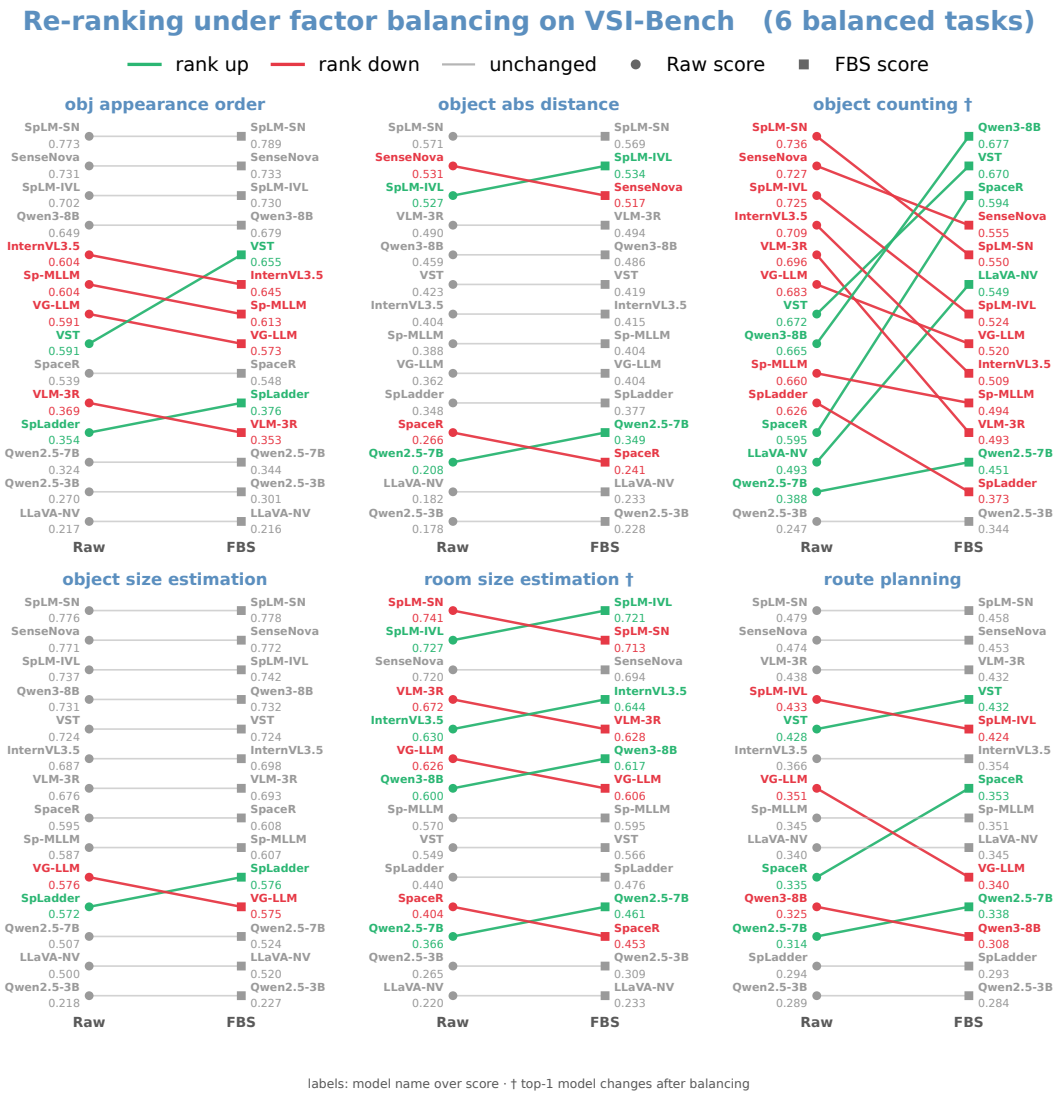

Figure G.2: **Re-ranking across balanced VSI-Bench tasks.** Left: observed scores and model ranks. Right: balanced scores and the resulting rank changes. Red indicates rank decreases, green indicates rank increases, and italics denote general-purpose models.

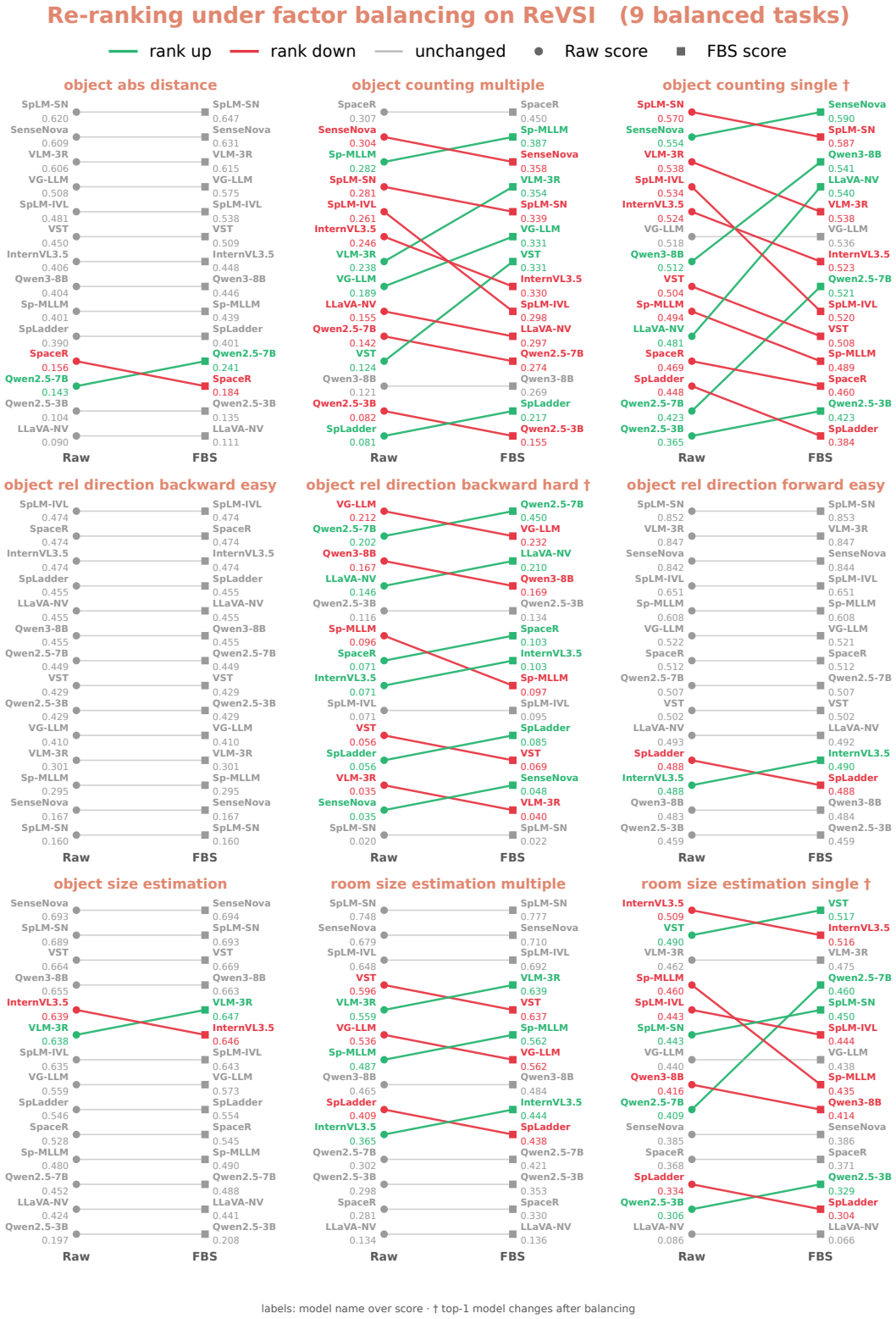

Figure G.3: **Re-ranking across balanced ReVSI tasks.** Conventions as in Fig. G.2.

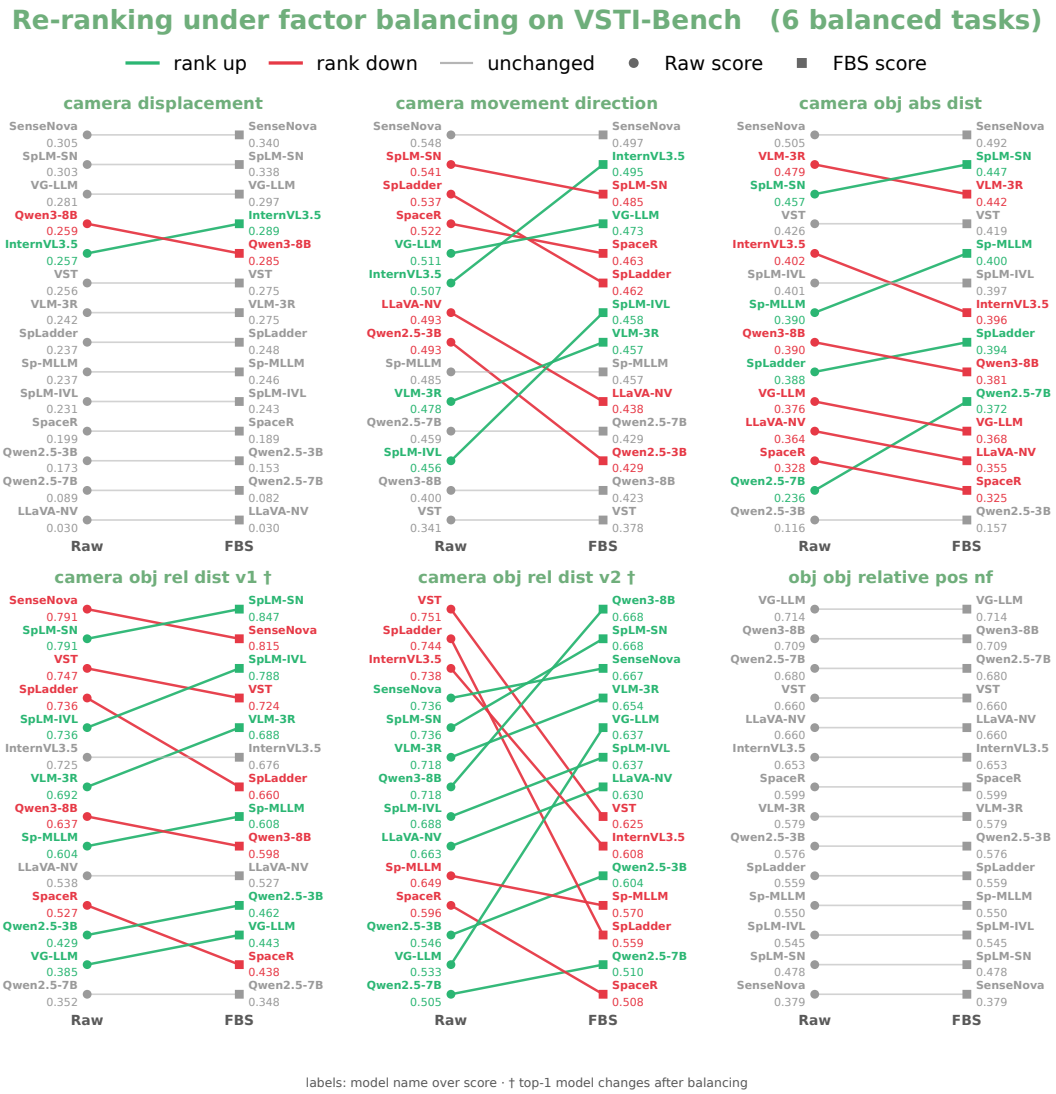

Figure G.4: **Re-ranking across balanced VSTI-Bench tasks.** Conventions as in Fig. G.2.



Figure G.5: **Re-ranking across balanced OSI-Bench tasks.** Conventions as in Fig. G.2.

Table G.3: **Ranking reversals by task** (14 models). We consider ordered model pairs (m, m') with $S_m^{\text{obs}} > S_{m'}^{\text{obs}}$ and count a reversal when the mean item score of m is lower than that of m' in at least half of the ground-truth bins containing at least 30 items. The final column reports, for each task, the reversed pair with the largest aggregate score gap and the number of eligible bins in which the aggregate leader trails.

Bench.	Task	GT bins	Reversed / ordered pairs	Share	Largest reversed pair (gap; bins lost)
VSI-Bench	Appearance order	4	6/89	7%	SpatialLadder-3B > LLaVA-NeXT-Video-7B (+0.138; 2/4)
	Abs. distance	4	13/91	14%	SpaceR-7B > Qwen2.5-VL-3B (+0.087; 2/4)
	Object counting	3	6/91	7%	SpatialLadder-3B > LLaVA-NeXT-Video-7B (+0.132; 2/3)
	Rel. direction	4	21/91	23%	Spatial-MLLM-5B > SpaceR-7B (+0.128; 2/4)
	Rel. distance	4	10/90	11%	LLaVA-NeXT-Video-7B > Qwen2.5-VL-7B (+0.054; 2/4)
	Object size	4	14/91	15%	Spatial-MLLM-5B > LLaVA-NeXT-Video-7B (+0.088; 2/4)
	Room size	4	7/91	8%	VST-7B > SpatialLadder-3B (+0.109; 2/4)
	Route planning	3	16/91	18%	VLM-3R-7B > InternVL3.5-8B (+0.072; 2/3)
ReVSI	Abs. distance	4	10/91	11%	SpatioLM-InternVL > SpatialLadder-3B (+0.091; 2/4)
	Counting (multi)	3	0/91	0%	–
	Counting (single)	4	16/91	18%	SpatialLadder-3B > Qwen2.5-VL-3B (+0.084; 2/4)
	Rel. dir. (bwd, easy)	3	17/84	20%	SpaceR-7B > VG-LLM-8B (+0.064; 2/3)
	Rel. dir. (bwd, hard)	4	5/86	6%	Qwen2.5-VL-3B > SpaceR-7B (+0.045; 2/4)
	Rel. dir. (fwd, easy)	3	11/91	12%	Spatial-MLLM-5B > Qwen2.5-VL-7B (+0.094; 2/3)
	Rel. dir. (fwd, hard)	4	18/90	20%	Spatial-MLLM-5B > Qwen3-VL-8B (+0.110; 2/4)
	Rel. dist. (closest)	4	9/90	10%	VST-7B > SpatialLadder-3B (+0.070; 2/4)
	Rel. dist. (farthest)	4	16/91	18%	VST-7B > Spatial-MLLM-5B (+0.138; 2/4)
	Object size	4	16/91	18%	SpatialLadder-3B > LLaVA-NeXT-Video-7B (+0.122; 2/4)
	Room size (multi)	2	29/91	32%	(fewer than 3 bins)
	Room size (single)	4	26/91	29%	Qwen2.5-VL-7B > SpatialLadder-3B (+0.075; 2/4)
	Route planning	4	33/90	37%	SpatioLM-SenseNova > Qwen3-VL-8B (+0.205; 2/4)
VSTI-Bench	Camera displacement	4	17/91	19%	InternVL3.5-8B > SpaceR-7B (+0.058; 2/4)
	Camera move dir.	4	74/90	82%	SpatioLM-SenseNova > VST-7B (+0.210; 2/4)
	Cam-obj abs. dist.	4	43/91	47%	SenseNova-SI-8B > SpaceR-7B (+0.177; 2/4)
	Cam-obj rel. dist. v1	1	6/89	7%	(fewer than 3 bins)
	Cam-obj rel. dist. v2	3	9/89	10%	Qwen3-VL-8B > Spatial-MLLM-5B (+0.069; 2/3)
	Cam-obj rel. dist. v3	9	15/90	17%	InternVL3.5-8B > VG-LLM-8B (+0.140; 5/9)
	Obj-obj rel. pos. (l/r)	2	45/91	49%	(fewer than 3 bins)
	Obj-obj rel. pos. (n/f)	2	70/90	78%	(fewer than 3 bins)
	Obj-obj rel. pos. (u/d)	2	43/91	47%	(fewer than 3 bins)
OSI-Bench	Abs. displacement	4	25/91	27%	LLaVA-NeXT-Video-7B > VG-LLM-8B (+0.066; 2/4)
	Abs. distance	4	20/91	22%	InternVL3.5-8B > SpaceR-7B (+0.074; 2/4)
	Abs. speed	4	57/91	63%	Qwen3-VL-8B > SpatialLadder-3B (+0.245; 2/4)
	Depth-aware counting	4	41/91	45%	SpatioLM-InternVL > Spatial-MLLM-5B (+0.285; 2/4)
	3D localization	4	17/91	19%	VG-LLM-8B > SpatioLM-InternVL (+0.075; 3/4)
	Rel. direction	4	57/88	65%	VST-7B > Qwen2.5-VL-7B (+0.146; 2/4)
	Rel. distance	4	30/90	33%	Qwen2.5-VL-3B > LLaVA-NeXT-Video-7B (+0.042; 2/4)
	Qualitative ego-motion	3	18/91	20%	VST-7B > VLM-3R-7B (+0.050; 2/3)
	Quantitative ego-motion	4	24/91	26%	Qwen3-VL-8B > VG-LLM-8B (+0.165; 2/4)

Additional analysis: Ranking reversals. Tab. G.3 reports, for each task, the fraction of ordered model pairs whose ranking is reversed in most ground-truth bins, together with the reversal having the largest aggregate score margin. We use the four tail-aware ground-truth bins defined in Sec. D (answer classes for multiple-choice tasks) and retain only bins containing at least 30 items. For a model pair (m, m') satisfying $S_m^{\text{obs}} > S_{m'}^{\text{obs}}$, we count a ranking reversal when the mean item score of m is lower than that of m' in at least half of the retained bins. Across the 34 tasks with at least three eligible bins, 717 of 3,068 ordered model pairs (23%) exhibit such reversals. These reversals are not limited to models with similar aggregate scores. On VSI-Bench's counting, SpatialLadder-3B leads LLaVA-NeXT-Video-7B by 0.13 in S_m^{obs} yet trails it in both bins containing three or more objects. On VSTI-Bench camera-object absolute distance, SenseNova-SI leads SpaceR by 0.18 overall while trailing in two of four distance bins. On OSI-Bench depth-aware counting, SpatioLM-InternVL leads Spatial-MLLM by 0.29 overall while trailing in two of four count bins.

Discretization sensitivity. Continuous factors enter CST and FBS as categorical variables, cut into four bins by the tail-aware rule of Sec. D: the top 20% of the sample forms the last bin and the value range of the remaining 80% is split into three equal-width bins. To test whether this choice drives the results we re-run the full pipeline, the CST diagnosis and the FBS balancing, under (A) $K = 3, \ldots, 10$ bins with the same rule generalized to $K - 1$ equal-width bins below the tail, and (B) three alternative cutting rules at $K = 4$: equal-frequency quartiles, one-dimensional k-means boundaries, and equal-width bins over the full value range. Only the edge function is swapped; the de-duplicated candidate set, the top-10 selection, c, the ratio band, the feasible target, and δ are those of the paper. Each setting is compared with the current one on the share of the current direct (task, factor) slots that remain direct, the Kendall τ between its balanced ranking and the current one (over tasks balanced in both settings), and the share of empty joint cells that the balancing has to work on.

Tabs. G.4 and G.5 show that the diagnosis is insensitive to the number of bins: raising K to 6, 8, or 10 keeps 88%, 83%, and 76% of the current direct slots and leaves the balanced ranking essentially unchanged ($\tau = 0.87$–0.91), while $K = 3$ is too coarse and loses 18% of the slots. Finer bins only cost support, as the share of empty joint cells rises from 28% at $K = 4$ to 48% at $K = 10$ without changing what is balanced, so $K = 4$ is the coarsest discretization that preserves the diagnosis. Any rule that spreads the sample evenly gives the same answer: equal-frequency and k-means bins retain 82% and 90% of the direct slots ($\tau = 0.89$ and 0.85). Equal-width bins fail (71% retained, $\tau = 0.73$) because on long-tailed factors most items fall into the first bin and 44% of the joint cells are empty; the tail-aware rule matches the equal-frequency diagnosis while keeping the upper tail as its own bin.

Table G.4: **Number of bins.** Full pipeline re-run with K tail-aware bins ($c = 0.02$, 14 models, 39 tasks). *Retained*: share of the current direct slots that remain direct. τ: Kendall's τ between the balanced ranking under each setting and the current setting, averaged over tasks balanced under both. *Empty cells*: share of the joint factor grid without observations, averaged over balanced tasks.

Bins K	Direct slots	Retained	Kendall's τ	Empty cells
$K = 3$	64	82%	0.90	18%
$K = 4$	72	100%	1.00	28%
$K = 6$	76	88%	0.90	40%
$K = 8$	75	83%	0.91	44%
$K = 10$	71	76%	0.87	48%

Table G.5: **Cutting rule at $K = 4$.** Full pipeline re-run using alternative discretization rules with $K = 4$ ($c = 0.02$, 14 models, 39 tasks). *Retained*: share of the current direct slots that remain direct. τ: Kendall's τ between the balanced ranking under each setting and the current setting, averaged over tasks balanced under both. *Empty cells*: share of the joint factor grid without observations, averaged over balanced tasks.

Rule ($K = 4$)	Direct slots	Retained	Kendall's τ	Empty cells
`q4_top20_eq3`	72	100%	1.00	28%
Quantile	65	82%	0.89	12%
k-means	73	90%	0.85	25%
Equal width	82	71%	0.73	44%

Cross-benchmark comparisons and common support. Several task families appear across multiple benchmarks, including counting in VSI-Bench, ReVSI, and OSI-Bench, and absolute distance in all four benchmarks. In an earlier study with nine models, we examined whether balancing each benchmark toward a frequency-neutral composition increases agreement between leaderboards for the same task family. The results highlight an important distinction between differences in *composition* and differences in *support*. When two benchmarks cover similar value ranges, their observed leaderboards already show substantial agreement: Kendall's τ ranges from 0.79 to 0.86 between VSI-Bench and ReVSI for counting, distance, and size, and balancing largely preserves this agreement. In contrast, when benchmarks cover different value ranges, leaderboard agreement can be weak or even negative. For indoor benchmarks versus OSI-Bench on absolute distance, for example, $\tau = -0.57$; restricting both benchmarks to their common value range increases it to $+0.07$, whereas equalizing bin frequencies within each benchmark has a much smaller effect. Thus, uniformizing each benchmark over its own bins does not define a common evaluation composition when their supports differ. Cross-benchmark comparisons of balanced scores can therefore remain misleading unless the evaluation support is aligned first. Our main analysis avoids this issue by comparing observed and balanced scores within each benchmark, where the item support is fixed; accordingly, we report FBS separately for each benchmark rather than pooling scores across benchmarks.

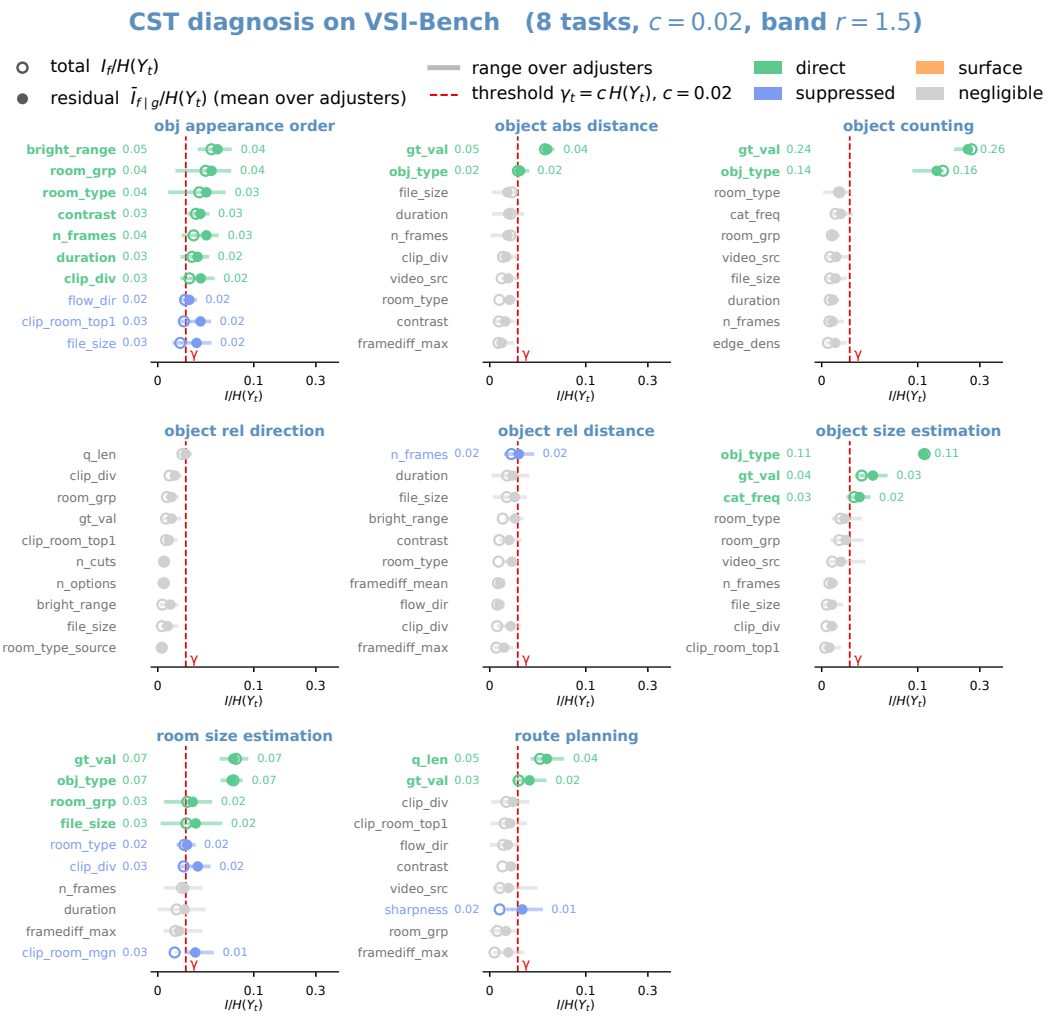

Figure G.6: **Full CST diagnosis on VSI-Bench (Yang et al., 2025a).** Extending Fig. 3, we show the top-10 task-agnostic factors for every VSI-Bench task. Hollow and filled markers denote total and residual dependence, respectively. Horizontal bands show the range across adjustment factors, and the dashed line marks the threshold γ_t. The x-axis uses an *asinh* scale for readability. Factors are colored by their CST regime: direct contributor, suppressed, surface correlate, or negligible.

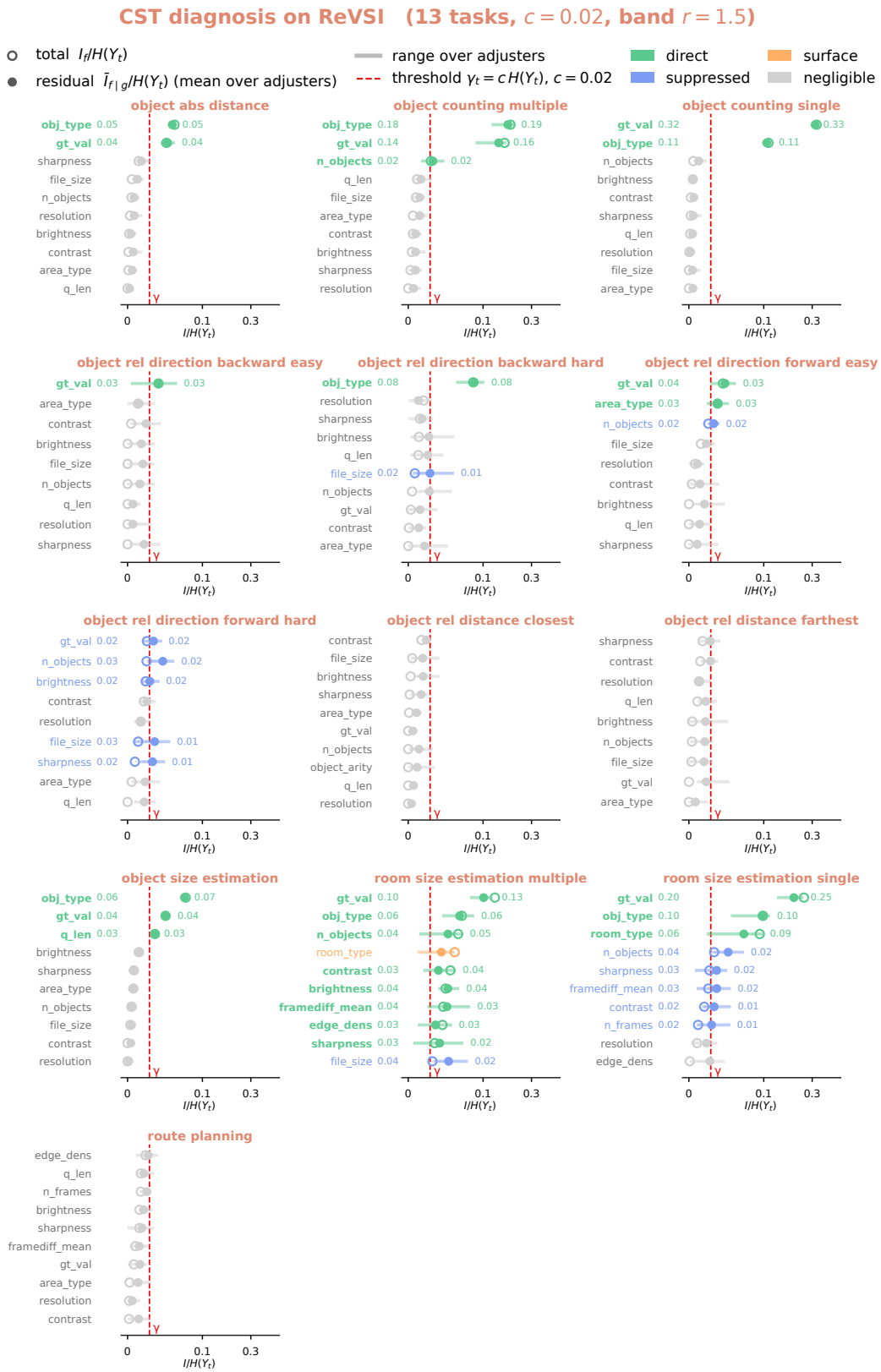

Figure G.7: **Full CST diagnosis on ReVSI (Zhang et al., 2026b).** Conventions as in Fig. G.6.

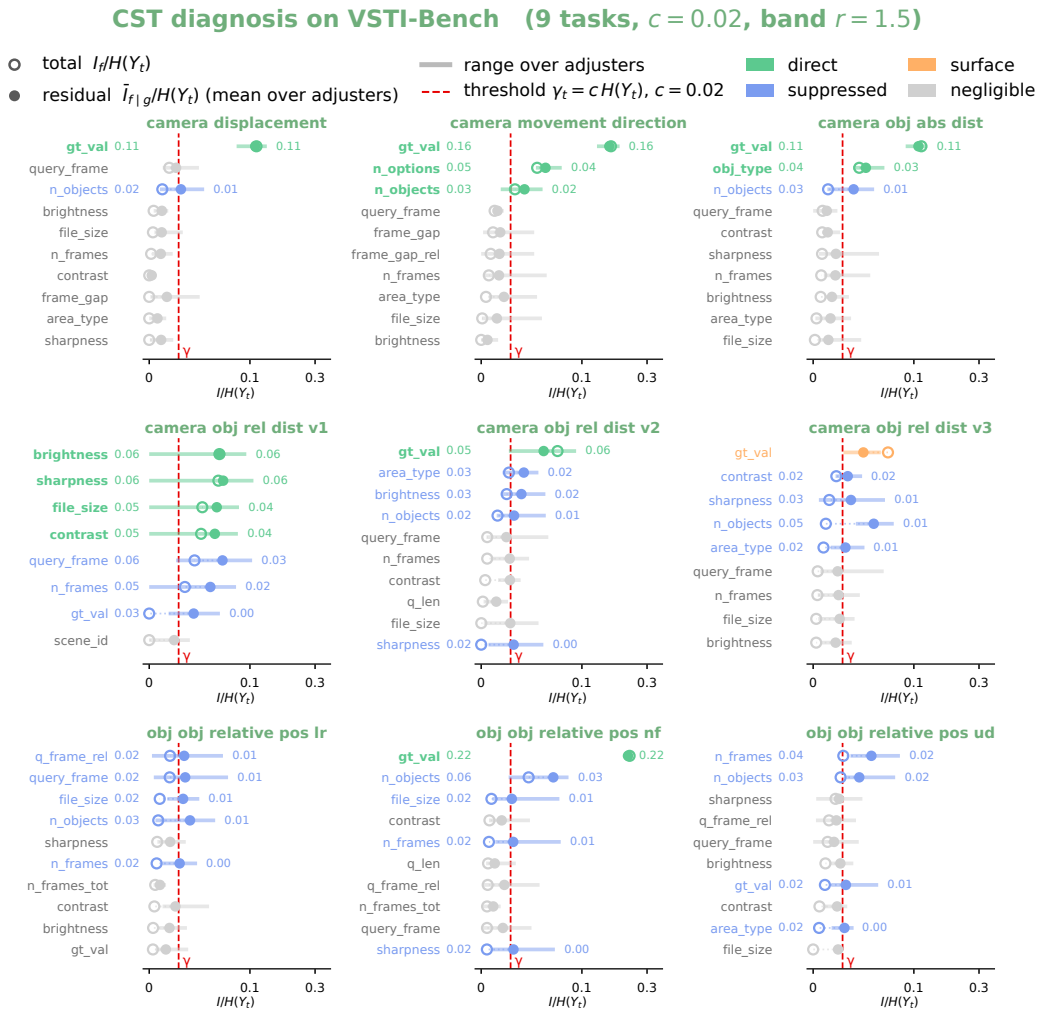

Figure G.8: **Full CST diagnosis on VSTI-Bench (Fan et al., 2026).** Conventions as in Fig. G.6.

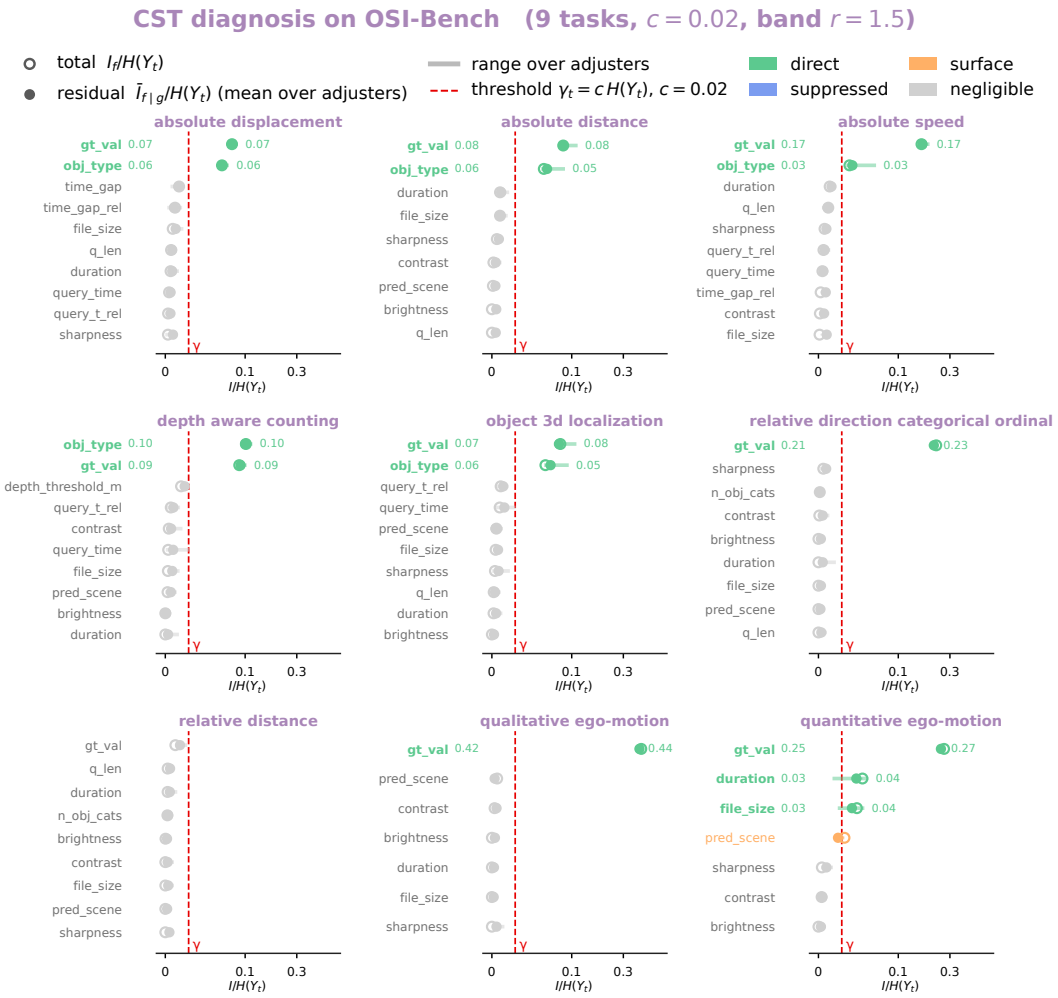

Figure G.9: **Full CST diagnosis on OSI-Bench (Wu et al., 2026d).** Conventions as in Fig. G.6.