Hidden Sensitivity in Spatial Reasoning Evaluation: Diagnosis and Re-ranking with VSI-Bench

Phillip Y. Lee [1] **Jin Yoo** [1] **Minseo Kim** [1] **Leonidas Guibas** [2] **Minhyuk Sung** [1]

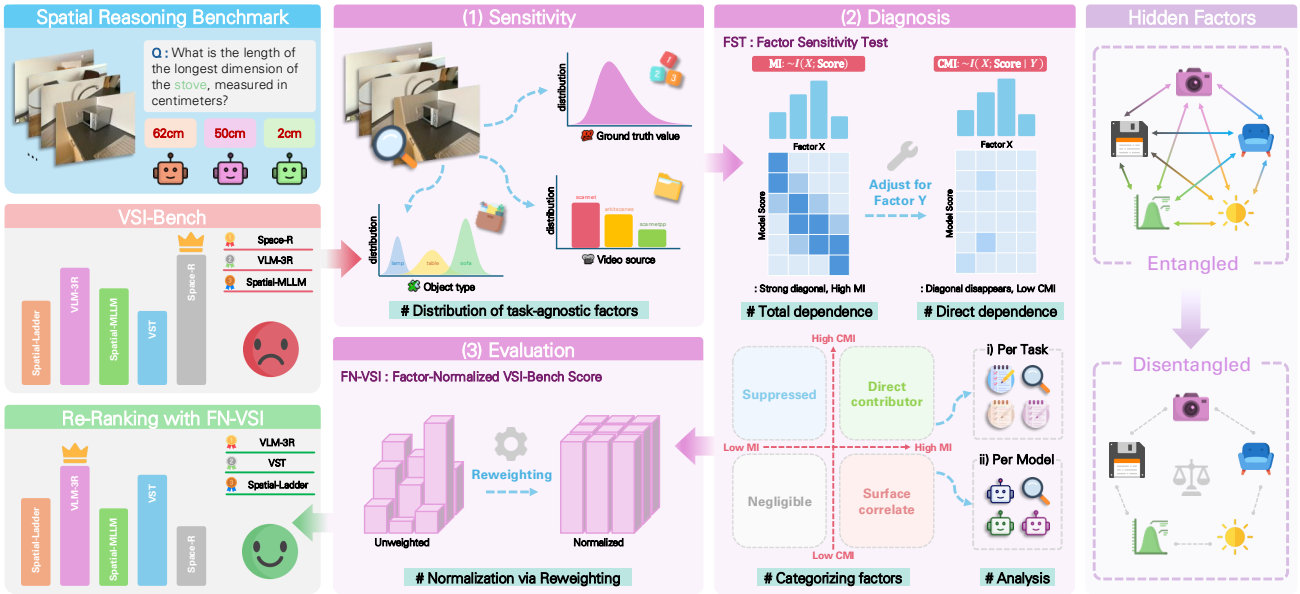

Figure 1. **Hidden sensitivity in spatial reasoning evaluation.** VSI-Bench (Yang et al., 2025a) scores are shaped not only by spatial reasoning ability, but also by hidden distributions of task-agnostic factors. We diagnose these entangled factor effects with FST and reweight the benchmark to obtain factor-normalized rankings that more faithfully reflect spatial reasoning capability.

Abstract

We introduce FN-VSI, a factor-normalized score for VSI-Bench based on an information-theoretic diagnostic, which substantially re-ranks vision-language models (VLMs). Spatial reasoning benchmarks for VLMs are typically reported as aggregate task scores, but such scores are not determined by model ability alone. VSI-Bench performance is strongly affected by *task-agnostic factors* such as video source, object category, ground-truth answer value, and low-level visual properties of the input video, none of which are the intended targets of evaluation. Since these factors are imbalanced and mutually entangled, raw score differences conflate spatial reasoning ability with benchmark composition. To disentangle their effects, we introduce FST, a diagnostic that estimates each factor's contribution by comparing its marginal and conditional mutual information with model scores, classifying factors as direct contributors, surface correlates, suppressed effects, or negligible factors. Across VSI-Bench tasks and multiple VLMs, ground-truth answer value and queried object type emerge as strong direct contributors, while several apparent sensitivities disappear after adjustment. Building on this diagnosis, FN-VSI reweights benchmark instances to neutralize the direct contributors identified by FST, indicating that reported spatial reasoning gains can depend on the factor mixture of the benchmark rather than on genuine ability differences.

[1]KAIST [2]Stanford University. Correspondence to: Phillip Y. Lee <phillip0701@kaist.ac.kr>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Spatial reasoning—counting objects in a scene, estimating their sizes and pairwise distances, and judging scene-scale

geometry such as room dimensions—is a foundational visual capability and an increasingly load-bearing claim in the modern vision–language model (VLM) literature, particularly as VLMs become central components of embodied agents. As datasets and models continue to scale, recent results suggest the emergence of non-trivial 3D and spatial competencies. Benchmarks such as VSI-Bench (Yang et al., 2025a) have therefore become de facto instruments for measuring these capabilities and steering subsequent research. Whether a model can be said to "reason spatially" is now determined largely by its performance on such leaderboards, which makes the construction of these benchmarks more consequential than they are usually treated as being.

Building such a benchmark entails many practical design choices: which scene datasets to draw from, which object categories and room types to include, and what range of answer values to query, and many others. These choices are typically treated as implementation details, fixed at construction time and rarely revisited during evaluation. We argue that these choices should not only be considered during benchmark construction, but also explicitly accounted for during evaluation. They constitute *task-agnostic factors*: properties of the evaluation data to which an ideal spatial reasoner should be invariant, yet which can nonetheless influence the scores reported by a benchmark.

Our work is motivated by two empirical observations. First, the distributions of task-agnostic factors in current spatial reasoning benchmarks are highly skewed: certain input video sources, object categories, and answer ranges are substantially over-represented. On VSI-Bench's object counting task, for instance, ground-truth answers are concentrated in a narrow low-value range. Re-sampling the evaluation set toward larger ground-truth values—while holding all other factors fixed—can drop the previously top-ranked VLM-3R (Fan et al., 2025) to third place. Second, measured performance is tightly coupled to that imbalance, so reported gaps between VLMs are partly attributable to which factor configurations *happen to* dominate the benchmark rather than to genuine differences in spatial reasoning. Until these two contributions are disentangled, leaderboard scores conflate model ability with benchmark composition, leaving composition-invariant spatial reasoning ability unobserved.

We propose an evaluation framework that (i) diagnoses which task-agnostic factors genuinely shape a VLM's measured score and (ii) produces a factor-normalized benchmark score that is robust to construction-time imbalances. Given the list of task-agnostic factors we consider, the framework consists of two components.

(1) Factor Sensitivity Test (FST). A naive per-factor analysis can be misleading because task-agnostic factors are often mutually entangled. For example, input video source may co-vary with object category, and object category may co-vary with answer range. As a result, a factor that appears strongly correlated with model performance may simply inherit the influence of another factor with which it is correlated. FST is an information-theoretic test that disentangles these effects. For each factor, it compares the *marginal* mutual information (Cover, 1999; Shannon, 1948) between the factor and the score with the *conditional* mutual information that remains when conditioned on another adjustment factor. Paired with a magnitude criterion on the conditional dependence, this ratio classifies each factor as a direct contributor, surface correlate, suppressed effect, or negligible factor, thereby distinguishing sensitivities that directly drive performance from those that arise from entanglement.

(2) Factor-Normalized VSI-Bench Score (FN-VSI). Given the genuine contributors flagged by FST, we define a benchmark-level score in which their joint distribution is reweighted to a canonical reference distribution, implemented through inverse-probability weighting of benchmark items (Rosenbaum & Rubin, 1983; Hernán & Robins, 2010). FN-VSI estimates the score a model would obtain on a hypothetical factor-balanced version of VSI-Bench, separating spatial reasoning ability from benchmark composition.

Factor normalization substantially reshapes the leaderboard (Fig. 6). On object counting, VG-LLM (Zheng et al., 2025) drops from second to fifth place under FN-VSI, with its score falling from 0.683 to 0.393 after its dependence on the direct contributoring factors is neutralized. This places it below its own base VLM (*i.e.*, Qwen2.5-VL-7B), suggesting that the apparent gain in counting ability may instead reflect overfitting to the benchmark's answer-value and target-object composition. On room size estimation, VG-LLM and VLM-3R swap the top two positions after normalization, indicating that the original ranking was partly induced by video source composition rather than by a stable performance advantage. We argue that these normalized rankings, rather than the raw rankings, provide the appropriate signal for evaluating today's VLMs and guiding their continued development.

In summary, this paper contributes:

- A taxonomy of ten task-agnostic factors in spatial reasoning benchmarks, instantiated on VSI-Bench, along with empirical evidence that small perturbations to these factors can shift model rankings at scales comparable to the gaps between competing models (Sec. 4).
- **FST**, an information-theoretic diagnostic that isolates each factor's direct effect on the score under mutual entanglement by comparing the marginal and conditional mutual information of the factor with the score, conditioned on an adjustment factor (Sec. 5).
- **FN-VSI**, a factor-normalized benchmark score, together with an empirical study showing that current VSI-Bench

rankings change substantively under FN-VSI and that several widely cited performance gaps are artifacts of benchmark composition (Sec. 6).

2. Related Work

2.1. Spatial Reasoning in Vision-Language Models

As vision-language models (VLMs) increasingly serve as the perception-and-reasoning backbone of physical embodied agents (Black et al., 2024; Lee et al., 2025; Koo et al., 2025; Team et al., 2025), assessing and improving their spatial reasoning ability has emerged as a critical challenge in the recent literature. Two complementary lines of work target this gap on the model side. The first is *data-centric*: SpatialVLM (Chen et al., 2024) showed that scalable spatial VQA synthesis can substantially improve spatial reasoning, with subsequent work (Yang et al., 2025c; Zhang et al., 2025; Daxberger et al., 2025; Chen et al., 2026; Li et al., 2025; AI et al., 2025; Yin et al., 2026; Lian et al., 2025; Liu et al., 2026; Brown et al., 2025a; Deng et al., 2025) extending this approach to richer tasks and stronger backbones. The second is *architectural*: SpatialRGPT (Cheng et al., 2024), Cambrian-1 (Tong et al., 2024) and related models (Zhao et al., 2025b; Chen et al., 2025; Zhang et al., 2026a; Fan et al., 2025; Zheng et al., 2025; Wu et al., 2025a) augment visual encoders with geometric inputs—such as depth maps, segmentation masks, or explicit 3D representations—to ground spatial questions beyond raw pixels.

A parallel body of work has developed benchmarks for measuring and training VLM spatial reasoning, with VSI-Bench (Yang et al., 2025a), CV-Bench (Tong et al., 2024), and RoboSpatial (Song et al., 2025) as representative efforts covering metric estimation, relation reasoning embodied grounding, extending coverage to navigation and long-horizon video reasoning (Wu et al., 2025c;b; Lin et al., 2025b; Yang et al., 2026; 2025b; Yu et al., 2025; Lin et al., 2025a). While these benchmarks and data suites have helped the community establish clearer targets, what their aggregate scores *actually reveal* about a model's spatial reasoning capability has received comparatively little attention. A recently growing line of work has begun to examine data shortcuts and contamination in spatial reasoning benchmarks (Brown et al., 2025b; Feng et al., 2025); we discuss this line of work in the following section.

2.2. Benchmark and Evaluation Validity

In the era of foundation models with near-human-level capabilities (Comanici et al., 2025; Anthropic, 2026; Hurst et al., 2024), faithfully benchmarking their abilities along specific dimensions has become more important than ever. A growing body of work questions whether existing benchmarks truly measure the capabilities they claim to assess (Neuhaus et al., 2026). Singh *et al*. (Singh et al., 2025) show how arena rankings can shift with unintentional factors, while VQA-CP (Agrawal et al., 2018) reveals language priors driving VQA performance. Zhao *et al*. (Zhao et al., 2025a) argue that the chain-of-thought (CoT) reasoning abilities of LLMs may stem from in-distribution data rather than genuine general reasoning. Gorla *et al*. (Gorla & Puduppully, 2026) traces apparent gains on tabular benchmarks to memorization rather than generalization. Similar concerns recently surface in visual and spatial reasoning: Mirage (Asadi et al., 2026) and VPBench (Feng et al., 2025) show that high benchmarks scores often rely on superficial data cues, and Ma *et al*. (Ma et al., 2026) further report that strong 3D-LLM performance can be reproduced without actual 3D understanding. Notably, Brown *et al*. (Brown et al., 2025b) reveal non-visual shortcuts in VSI-Bench (Yang et al., 2025a) and propose a fairer pruned subset, while a concurrent work ReVSI (Zhang et al., 2026b) identifies data incompleteness in VSI-Bench and introduces a more carefully curated version. Motivated by these works, we take a statistically grounded approach to diagnosing under-recognized yet critical sensitivity factors in the benchmark, providing a systematic tool for assessing how much of the reported gain is attributable to such factors.

2.3. Statistical Disentanglement of Confounded Effects

Our framework builds on a long line of work on disentangling the effects of correlated factors in observational data. Information-theoretic dependence measures provide a natural foundation: mutual information (Cover, 1999; Shannon, 1948) quantifies the overall association between a factor and an outcome, while conditional mutual information measures the association that remains after conditioning on another factor. Related ideas underlie a broad family of statistical and causal adjustment methods, including causal mediation analysis and natural direct effects (VanderWeele, 2015), propensity-score methods and inverse-probability weighting (Rosenbaum & Rubin, 1983; Hernán & Robins, 2010), and the g-formula (Robins, 1986). Closer to our setting, the visual question answering (VQA) community has studied language-prior bias through a causal lens (Niu et al., 2021; Cadene et al., 2019; Clark et al., 2019), and recent work extends such structural-causal and counterfactual analyses to vision-language models (VLMs) for diagnosing spurious correlations and reasoning failures (Xu et al., 2026; Komanduri et al., 2025; Yang et al., 2025e). Our work extends these efforts to spatial reasoning benchmarks. Specifically, we identify hidden task-agnostic sensitivities in VSI-Bench (Yang et al., 2025a), formalize them using mutual-information-based diagnostics (Cover, 1999; Shannon, 1948) to quantify which factors directly influence the score, and construct a factor-normalized re-evaluation protocol that adjusts benchmark scores for these sensitive factors via inverse-probability weighting (Rosenbaum & Rubin, 1983; Hernán & Robins, 2010; Horvitz & Thompson, 1952).

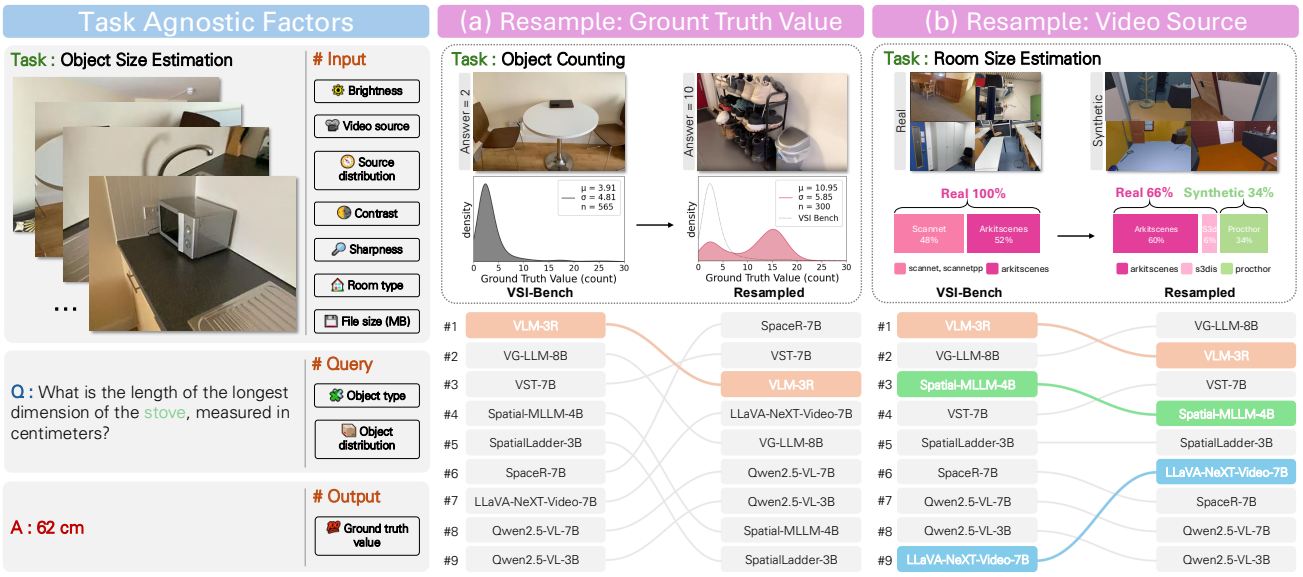

Figure 2. **Task-agnostic factors.** While VSI-Bench (Yang et al., 2025a) targets spatial reasoning evaluation, its scores can also be influenced by multiple *task-agnostic* factors. Notably, simply re-sampling the evaluation data to shift the distributions of (a) ground-truth values or (b) video sources can substantially perturb the model leaderboard. Details are provided in Sec. 4

3. Overview

We aim to construct a *factor-normalized* VSI-Bench score: a metric that reweights each task-agnostic factor configuration toward a reference distribution, such that cross-model comparisons reflect genuine spatial reasoning ability rather than the specific factor configuration under which each instance is evaluated. Given the task-agnostic factors we consider (Sec. 4), we develop this score in three steps. **(1) Sensitivity (Sec. 4).** We identify task-agnostic factors of VSI-Bench (Yang et al., 2025a) and show that small perturbations to these factors can shift per-task accuracy by margins comparable to the gaps between competing models. **(2) Diagnosis (Sec. 5).** Because these factors are mutually entangled, we introduce the Factor Sensitivity Test (FST), an information-theoretic estimator that compares each factor's *marginal* mutual information with the score to its *conditional* mutual information given an adjustment factor. We apply it across VSI-Bench to isolate the direct, genuine drivers of score variation. **(3) Evaluation (Sec. 6).** Given the dominant unintended factorsn identified by FST, we derive a factor-normalized score, FN-VSI, via inverse probability weighting. We then use FN-VSI to re-rank VLMs and compare the resulting rankings against the standard aggregate.

4. Hidden Sensitivity in VSI-Bench

Task-agnostic factors in VSI-Bench. We catalog ten task-agnostic factors in VSI-Bench (Yang et al., 2025a), covering three core aspects of the evaluation data: *input*, which captures properties of the video itself; *query*, which captures properties of the objects referenced in the question; and *output*, which captures properties of the answer to be produced. These factors range from simple metadata, such as ground-truth value and file size, to higher-level semantic descriptors, such as room type. We list the full set in Appendix A.2, along with per-factor definitions and value sets.

Sensitivity to task-agnostic factors. None of these factors directly describes the *spatial reasoning* in a question, yet each can visibly shape the final score. Fig. 2-(b) illustrates this effect through two notable changes in the model rankings. **(a) Ground-truth resampling.** On `object counting`, we re-sample a VSI-Bench-like evaluation set to skew the ground-truth answer distribution toward larger values, while leaving other factors unchanged (Appendix A.5). After this single change, VLM-3R (Fan et al., 2025), which ranks first under the original distribution, drops to third. **(b) Video-source resampling.** On `room size estimation`, we modestly shift the video-source mixture by replacing a fraction of real scenes with synthetic ones, which again alters the ranking. VLM-3R, the previous top-ranked model, and Spatial-MLLM (Wu et al., 2025a) each drop by one rank, while LLaVA-NeXT-Video (Zhang et al., 2024) rises from last place to the middle of the table.

In both perturbations, the spatial reasoning task under evaluation remains unchanged; only the distribution of task-agnostic factors is varied. The benchmark score is therefore not a property of model ability alone, but of model ability *under a particular factor distribution*. The leaderboard observed by practitioners is only one of many equally valid rankings that could have been reported under alternative,

yet reasonable, benchmark compositions. The remainder of the paper asks which factors drive these shifts, and what evaluation would look like if their influence were explicitly controlled.

5. Which Factors Actually Contribute?

While the observations in Sec. 4 show that small perturbations to task-agnostic factors can substantially move VSI-Bench (Yang et al., 2025a) rankings, they do not yet reveal which factors are *actually* responsible for these shifts. In practice, varying one factor often changes the distribution of others as well; we quantify this entanglement in Appendix A.3. We therefore turn the question around and ask which factors retain their influence after their entanglement with other factors is removed. Sec. 5.1 formalizes this disentanglement question as a measurement problem over the joint distribution of task-agnostic factors and benchmark scores. Sec. 5.2 introduces FST, an information-theoretic estimator for per-factor effects, and Sec. 5.3 applies FST across VSI-Bench to identify the direct contributors.

5.1. Problem Formulation

Let $\mathcal{F}$ denote the finite set of task-agnostic factors introduced in Sec. 4, with $|\mathcal{F}| = 10$. Each factor $f \in \mathcal{F}$ is represented by a categorical factor variable X_f taking values in a finite level set $\mathcal{V}_f$. Continuous metadata are discretized and high-cardinality categorical metadata are coarsened, so that $|\mathcal{V}_f|$ is typically between 2 and 4. Let S denote the per-instance score random variable (*i.e.*, MRA on VSI-Bench (Yang et al., 2025a)). FST considers one factor $f \in \mathcal{F}$ at a time as the *focal factor* and pairs it with a single *adjustment factor* $g \in \mathcal{F} \setminus \{f\}$. We then compare the total (or marginal) dependence between the score and the focal factor, $I(S; X_f)$, with the direct (or conditional) dependence $I(S; X_f \mid X_g)$, which measures how much dependence remains once the adjustment factor is held fixed.

Total and direct dependence estimation with mutual information. We need a measure that quantifies how strongly the score variable S depends on a focal factor variable X_f, and that admits a natural conditional analogue: how strongly S depends on X_f once an adjustment factor variable X_g is held fixed. *Mutual information* (MI) (Cover, 1999; Shannon, 1948) is well-suited to this purpose: it requires no parametric assumption about the relationship between S and X_f, captures arbitrary nonlinear and non-monotone dependencies, and applies directly to the finite categorical variables induced by our factor-level construction. Crucially, the *conditional mutual information* (CMI) $I(S; X_f \mid X_g) = 0$ holds if and only if $S \perp X_f \mid X_g$, giving an information-theoretic test for whether the apparent association between the score and the focal factor is explained by the adjustment factor X_g (Cover, 1999)—exactly our disentanglement question.

We therefore measure the relationship between each focal factor and the score in two complementary ways. The *total dependence* captures all associations between X_f and S, including those induced by entanglement with other factors. The *direct dependence* captures only the association that remains after holding the adjustment factor X_g fixed. Both are defined as population-level mutual information quantities:

$$I(S; X_f) = \sum_{s \in \mathcal{S}} \sum_{x_f \in \mathcal{V}_f} p(s, x_f) \log \frac{p(s, x_f)}{p(s)\, p(x_f)}, \quad (1)$$

$$I(S; X_f \mid X_g) = \sum_{x_g \in \mathcal{V}_g} p(x_g)\, I(S; X_f \mid X_g = x_g), \quad (2)$$

where the conditional term $I(S; X_f \mid X_g = x_g)$ is the mutual information between S and X_f evaluated within the stratum $X_g = x_g$. Thus, $I(S; X_f)$ measures the marginal dependence between the score and the focal factor, whereas $I(S; X_f \mid X_g)$ measures the dependence that remains after holding the adjustment factor fixed. The diagnostic question is how $I(S; X_f \mid X_g)$ changes relative to $I(S; X_f)$; the next section turns this comparison into a factor-sensitivity criterion.

5.2. FST: Factor Sensitivity Test

We now introduce FST, a diagnostic procedure that uses mutual information (Cover, 1999; Shannon, 1948) to identify which focal-factor dependencies persist after accounting for an adjustment factor. FST is designed to answer the following question: how much of the association between the score S and a focal factor X_f remains after holding an adjustment factor X_g fixed? To quantify this, we define the *sensitivity ratio* $\mathcal{R}_{f|g}$ as follows:

$$\mathcal{R}_{f|g} = \frac{I(S; X_f \mid X_g)}{I(S; X_f)} \in [0, \infty), \quad (3)$$

with the convention $\mathcal{R}_{f|g} := 0$ when $I(S; X_f) = 0$. The ratio measures the fraction of the focal factor's total dependence on the score that remains after conditioning on the adjustment factor.

From population quantities to finite-sample estimators. Both $I(S; X_f)$ and $I(S; X_f \mid X_g)$ are population-level quantities. In practice, we approximate them with plug-in estimates computed from empirical contingency tables, $\widehat{I}(S; X_f)$ and $\widehat{I}(S; X_f \mid X_g)$, followed by a Miller–Madow finite-sample bias correction (Miller, 1955). For CMI, we also apply a minimum stratum-size filter, which plays the role of a common-support restriction in inverse-probability weighting (Rosenbaum & Rubin, 1983; Horvitz & Thompson, 1952). All sensitivity quantities are computed from these empirical estimates, and we drop the hat notation when it is clear from context.

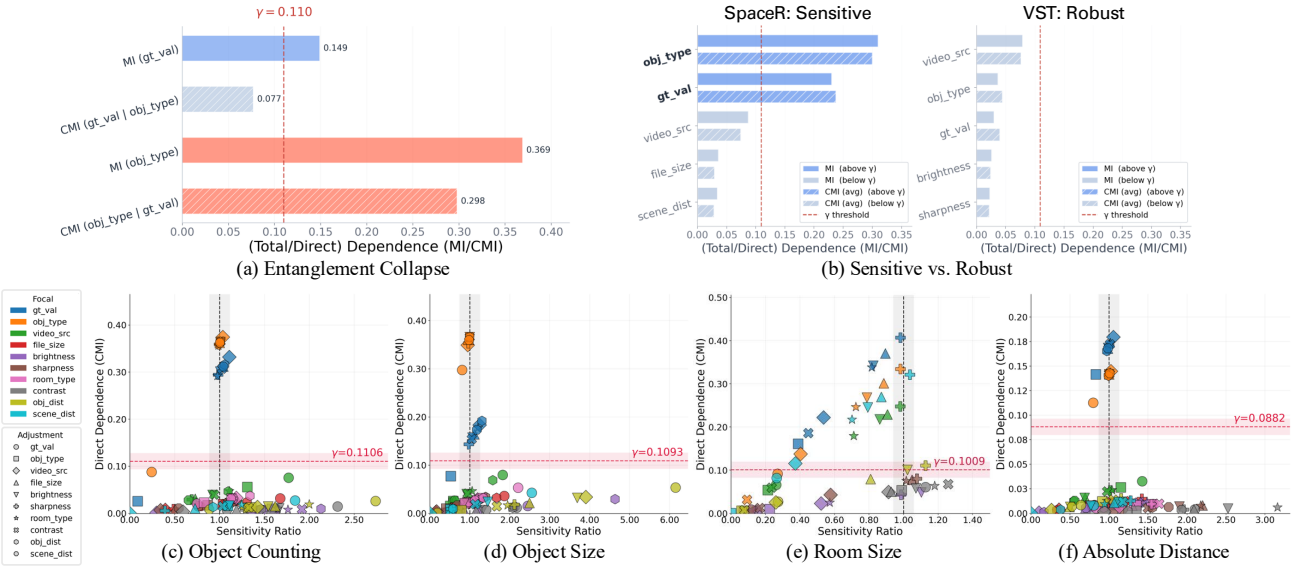

Figure 3. **FST Diagnosis.** Each plot presents a representative FST analysis on VSI-Bench (Yang et al., 2025a). (a) FST shows that a factor with high total dependence can become insignificant after adjustment. For example, after controlling for `obj_type`, the dependence between the score and `gt_val` falls below the threshold γ, indicating that `gt_val` may act as a surface correlate rather than a direct contributor in this setting. (b) FST further reveals model-level variation in robustness to task-agnostic factors. (c)–(f) plot the sensitivity ratio (Eq. 3) against conditional mutual information (CMI) for each task. Detailed explanations are provided in Sec. 5.3.

Categorizing factors. The sensitivity ratio $\mathcal{R}_{f|g}$ alone is insufficient for factor classification: it captures only the *relative* change from total to direct dependence, while leaving the *absolute* strength of the remaining dependence unconstrained. A factor whose total and direct dependence are both vanishingly small can yield $\mathcal{R}_{f|g} \approx 1$, just as a factor with genuinely strong dependence can, even though only the latter meaningfully shapes the score. We therefore pair the ratio with a magnitude criterion on $I(S; X_f \mid X_g)$, and assign each pair (f, g) to one of the four regimes summarized in Tab. 1. Because score-bin entropy differs across tasks, we use a task-adaptive magnitude threshold rather than a fixed MI cutoff. For each task τ, we set $\gamma_\tau = c \cdot H(S_\tau)$, where S_τ is the score-bin random variable restricted to task τ, $H(\cdot)$ denotes entropy, and $c = 0.05$. This normalizes the criterion by the task-specific score uncertainty in the score distribution, with c chosen to reflect a small-effect threshold in line with conventions for standardized effect-size reporting (Cohen, 2013; Lakens, 2013).

5.3. Diagnosis on VSI-Bench

We apply FST to VSI-Bench (Yang et al., 2025a) to identify, for each task, which task-agnostic factors directly contribute to score variation and which only *appear* influential due to entanglement with other factors.

Setup. We analyze four numerical tasks—`object counting`, `object size`, `room size`, and `absolute distance`—across nine VLMs, spanning general-purpose models and variants fine-tuned for spatial reasoning. Full model list is provided in Appendix A.5. For each task τ, we evaluate the ten task-agnostic factors introduced in Sec. 4. For conciseness, we use abbreviated factor identifiers throughout the following analysis; full list of factors and their identifiers is provided in Tab. 2. The raw per-instance MRA score is discretized into $B = 10$ equal-width bins of width 0.1 to obtain the score-bin variable S; high-cardinality categorical factors (*e.g.*, `obj_type`) are coarsened into 3–4 levels.

Data pool construction. VSI-Bench (Yang et al., 2025a) contains 2,640 numerical task instances drawn from three source datasets: ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ArkitScenes (Baruch et al., 2021). To support a more reliable FST diagnosis, we prepare a larger and more diverse data pool while preserving the task format as VSI-Bench. Specifically, we use VSI-590K (Yang et al., 2025d), a larger collection of video-based spatial reasoning examples generated through a similar pipeline and covering the same set of tasks. Because both datasets may share sources with the training data of evaluated VLMs, we rigorously filter out instances likely to cause contamination. The detailed filtering procedure is outlined in Appendix A.5 and results in a final suite of 42,836 samples.

Table 1. **Sensitivity regimes.**

Regime	$I(S; X_f \mid X_g)$	$\mathcal{R}_{f\|g}$
Direct contributor	$> \gamma$	≈ 1
Suppressed	$> \gamma$	$\gg 1$
Surface correlate	$\leq \gamma$	$\to 0$
Negligible	$\leq \gamma$	–

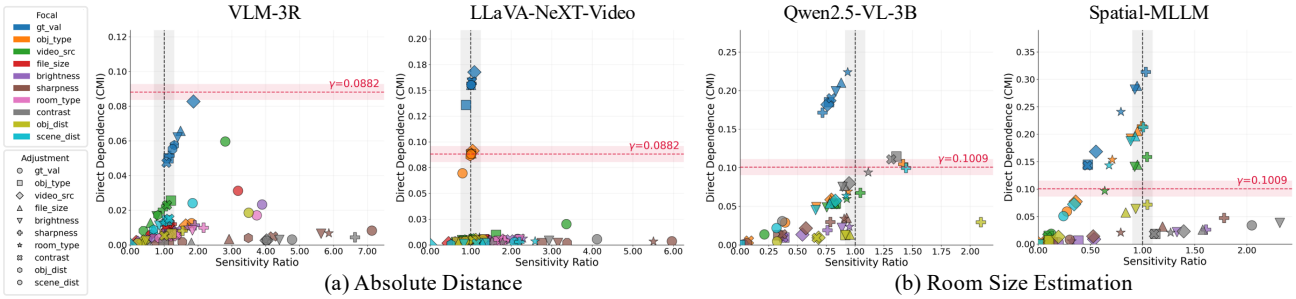

Figure 4. **Model-wise analysis with FST.** Each plot highlights representative model-wise factor-sensitivity diagnoses on VSI-Bench (Yang et al., 2025a) obtained with FST. Detailed descriptions of the results are provided in Sec. 5.3.

Key diagnosis with FST. We present representative FST analyses of factor sensitivity in VSI-Bench in Fig. 3.

Distinguishing surface and direct contributors. Fig. 3-(a) shows that FST separates factors that merely *appear* associated with the score through total dependence from those that remain as direct contributors after controlling other factors. On `object size` task, MI ranks `obj_type` first and `gt_val` second, with values of 0.369 and 0.149, respectively. However, after controlling the adjustment factor, only `obj_type` remains above the direct contributor threshold: its CMI remains high at 0.298, exceeding the task-wise threshold $\gamma_\tau = 0.110$, whereas `gt_val` drops to 0.070. This suggests that the apparent sensitivity to ground-truth answer value is partly explained by the queried object type, while object type itself remains a direct contributor.

Model-wise sensitivity. Fig. 3-(b) further reveals that the same benchmark factors can affect different models to substantially different degrees. On `object size` task, SpaceR (Ouyang et al., 2025) is highly factor-sensitive: two of the ten tested factors, `obj_type` and `gt_val`, remain direct contributors after disentanglement. Its strongest conditional dependence is on `obj_type`, with a CMI of 0.300, corresponding to 2.74× the task threshold γ_τ. In contrast, VST (Yang et al., 2025c) is comparatively factor-robust: none of its factors exceed the threshold, and even its largest CMI remains below γ_τ. This indicates that benchmark composition does not shift all model scores uniformly; instead, it can preferentially benefit models whose sensitivities align with the benchmark's factor distribution.

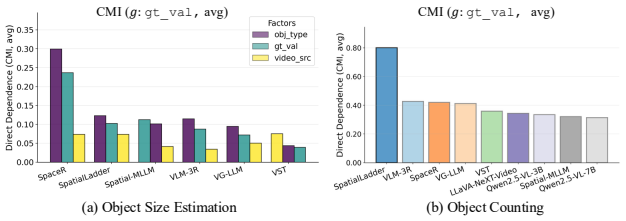

Figure 5. **Model-wise analysis with FST.** Each plot highlights representative model-wise factor-sensitivity diagnoses on VSI-Bench (Yang et al., 2025a) obtained with FST. Detailed descriptions of the results are provided in Sec. 5.3.

Task-wise diagnosis. Fig. 3-(c)–(f) summarize the FST diagnosis across all task-agnostic factors for the four numerical tasks in VSI-Bench. The set of direct contributing factors varies by task. `Room size` exhibits the broadest sensitivity profile, retaining four of the ten tested factors as direct contributors: `gt_val`, `obj_type`, `scene_dist`, and `file_size`. In contrast, `absolute distance` retains only two contributors, `gt_val` and `obj_type`. `Object counting` and `object size` also retain this recurring pair, although their relative strengths differ after adjustment. Overall, FST shows that VSI-Bench does not have a single universal source of sensitivity; rather, each task exhibits its own vulnerability profile, centered around a recurring answer-value/object-type axis.

Model-wise diagnosis. Fig. 4 presents a model-wise FST analysis, revealing notable variation in factor sensitivity across models. **(a)** On `absolute distance`, VLM-3R (left) is the only model for which no factor exceeds the direct-contributor threshold defined in Tab. 1, suggesting that its final benchmark score is not dominated by a specific task-agnostic factors. Other models, including LLaVA-NeXT-Video (right), exhibit direct dependence on one or more task-agnostic factors according to conditional mutual information (CMI). **(b)** Another notable pattern appears on `room size`. Qwen2.5-VL-3B (Bai et al., 2025) has only `gt_val` as a direct contributor to accuracy, whereas its variant Spatial-MLLM (Wu et al., 2025a) is additionally sensitive to `video_src` and `scene_dist`. This difference suggests that fine-tuning and integrating additional features (Wang et al., 2025) may introduce new unintended sensitivities to task-agnostic factors.

Fig. 5 shows additional model-wise sensitivity checks enabled by FST. **(a)** On `object size` task, `obj_type` emerges as the strongest direct contributor for nearly all evaluated models, consistent with the object-category shortcut identified by prior work (Brown et al., 2025b). **(b)** The bar plot reports results on `object counting`, ranking models by their average direct dependence on the remaining task-agnostic factors after adjusting for ground-truth value. Here, we fix $g =$ `gt_val` as the adjustment factor and average the CMI over focal factors f ranging over the

remaining task-agnostic factors. SpatialLadder (Li et al., 2025) ranks first with a mean value of over 0.8, nearly twice that of the second-highest model, VLM-3R. This indicates a disproportionate sensitivity to ground-truth values across different models.

6. FN-VSI: Factor-Normalized Score

In Sec. 5.3, FST diagnosed which factors genuinely drive each task's score and which merely *appear* influential due to entanglement. Building on these insights, we now shift our focus to evaluation. We propose FN-VSI, a *factor-normalized* score for VSI-Bench. By controlling for the influence of each task's dominant factors, this metric evaluates true spatial reasoning rather than a model's sensitivity to the benchmark's factor distribution. To achieve this, we employ a principle and intuitive adjustment mechanism: inverse-probability weighting (Rosenbaum & Rubin, 1983; Hernán & Robins, 2010).

Normalization via reweighting. All notation in this paragraph is local to a single task, and the same construction is applied independently for each task. Let $\mathcal{I}$ denote the set of task instances, and let $s_i \in [0, 1]$ be the score of a model on instance $i \in \mathcal{I}$. The standard aggregate score is the unweighted mean,

$$\bar{s} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_i. \tag{4}$$

Let $\mathcal{G} \subseteq \mathcal{F}$ be the set of *direct contributing factors* identified by FST. Their finite joint level space is

$$\mathcal{C}_{\mathcal{G}} = \prod_{f \in \mathcal{G}} \mathcal{V}_f. \tag{5}$$

For a cell $c \in \mathcal{C}_{\mathcal{G}}$, we write $i \in c$ when instance i matches the factor-level assignment specified by c. Let

$$n_c = |\{i \in \mathcal{I} : i \in c\}|, \qquad \widehat{\mathcal{C}}_{\mathcal{G}} = \{c \in \mathcal{C}_{\mathcal{G}} : n_c > 0\}. \tag{6}$$

The standard score $\bar{s}$ weights cells in proportion to their benchmark frequency. In contrast, we use a uniform target over the supported cells $\widehat{\mathcal{C}}_{\mathcal{G}}$, equivalently assigning each instance in cell c a weight proportional to $1/n_c$. The factor-normalized score is

$$s^* = \frac{1}{|\widehat{\mathcal{C}}_{\mathcal{G}}|} \sum_{c \in \widehat{\mathcal{C}}_{\mathcal{G}}} \frac{1}{n_c} \sum_{i \in \mathcal{I} : i \in c} s_i. \tag{7}$$

Thus, s^* corresponds to a uniform average of cell-wise mean scores: every supported configuration of the direct contributing factors contributes equally, independent of its frequency in the original benchmark. When $\mathcal{G} = \emptyset$, $\mathcal{C}_{\mathcal{G}}$ contains a single trivial cell, and Eq. 7 reduces to the original aggregate score $\bar{s}$. To avoid instability from empty cells or severe cell imbalance, which can over-weight sparse configurations, we apply filtering and thresholding as detailed in Appendix Sec. A.1.

Re-Ranking with FN-VSI. Finally, we re-rank the nine evaluated VLMs on each task by replacing the standard VSI-Bench aggregate score $\bar{s}$ with the factor-normalized score s^* defined in Eq. 7. Fig. 6 summarizes the resulting rankings. Across tasks, the model orderings induced by s^* differ substantially from those induced by $\bar{s}$, including notable rank flips among top-performing VLMs.

First, on `object counting`, VLM-3R drops from first to third place after factor normalization, with its score decreasing from 0.696 to 0.439 once the supported cells defined by the direct contributing factors, `gt_val` and `obj_type`, are controlled via reweighting. Under the factor-normalized score, VG-LLM even falls below its own base model, Qwen2.5-VL-7B. This suggests that the apparent gain in counting under $\bar{s}$ may instead reflect alignment with the benchmark's answer-value and target-object composition, rather than a genuine improvement in counting ability. Second, on `object size`, VST and VLM-3R remain the top models, but the remaining rankings change substantially after factor normalization. Spatial-MLLM rises from fourth to third place, whereas SpaceR drops from third to eighth, with its score falling from 0.595 to 0.475. LLaVA-NeXT-Video, by contrast, moves upward from eighth to sixth.

Third, on `room size`, controlling for `scene_dist` and `gt_val` changes the ordering of the leading models: the first and second models swap positions, as do the third and fourth models. Since `room size` contains more than two direct contributors identified by FST, we report re-rankings under additional factor combinations in Appendix (Fig. 11). Lastly, on `object absolute distance`, factor normalization produces substantial reordering across the leaderboard. VLM-3R drops from first to second place, with its score decreasing from 0.490 to 0.379, while VG-LLM rises from fourth to first, increasing from 0.362 to 0.397. These shifts indicate that the standard `object absolute distance` ranking is strongly affected by the joint distribution of `gt_val` and `obj_type`, and that factor normalization can reveal qualitatively different model comparisons even on the same task.

Score shift from VSI-Bench to FN-VSI. While Fig. 6 shows how factor normalization changes model rankings, Fig. 7 further reveals how much each model's measured accuracy changes when moving from VSI-Bench to FN-VSI. Across tasks, the score shifts are highly heterogeneous. On `object counting`, most top-ranked models experience large drops after normalization, while lower-ranked Qwen models improve, indicating that the original benchmark distribution favors some models more than others. `Room size` exhibits a similar pattern of accuracy drops: nearly all models decrease after normalization, although the magnitude of the drop varies substantially across models. The shifts are more varied on `object size` and

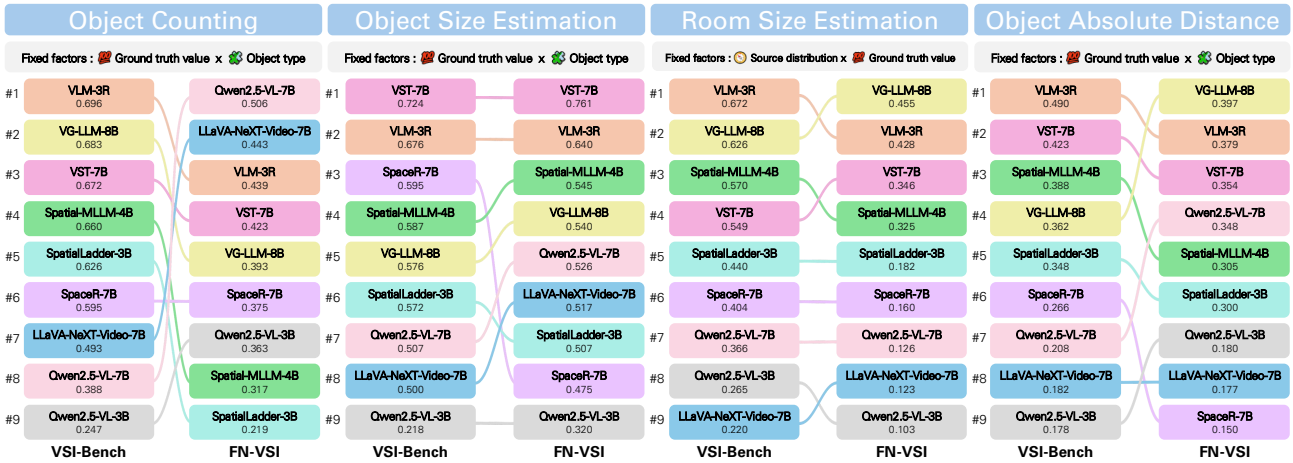

Figure 6. **Re-ranking with FN-VSI.** Factor-normalizing model accuracies using the key contributing factors identified by FST leads to notable rank shifts among VLMs, as detailed in Sec. 6.

`object absolute distance`. Some models remain relatively stable, while others either drop or improve after normalization. For instance, on `object absolute distance`, VG-LLM and Qwen2.5-VL-7B improve under FN-VSI, whereas VLM-3R, Spatial-MLLM, and SpaceR decrease. These model-specific trends show that factor normalization affects not only the ordering of models, but also the measured performance gaps between them.

These drops, swaps, and gap reversals suggest that the standard aggregate score can overstate model advantages when the evaluation distribution aligns with model-specific factor dependencies. Overall, the observed shifts show that standard VSI-Bench scores are sensitive to the distribution of task-agnostic factors. Controlling for the direct contributors identified by FST is therefore necessary for comparing models by their spatial reasoning ability, rather than by their alignment with the particular factor mixture present in the benchmark.

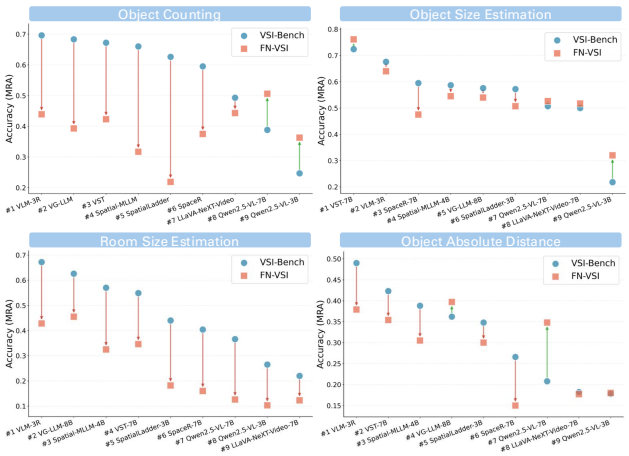

Figure 7. **Score shift from VSI-Bench to FN-VSI.** Each plot shows the model-wise scores before and after factor normalization. Arrows show the direction and magnitude of each model's score change after normalization. The heterogeneous shifts across models and tasks highlight that VSI-Bench scores are fragile to task-agnostic factor composition, significantly affecting the measured performance gaps.

7. Conclusion

We revisited what aggregate scores on VSI-Bench (Yang et al., 2025a) actually measure. We showed that task-agnostic factors in the evaluation set can shape per-task accuracy at a scale comparable to the gaps between competing models, and that small perturbations to these factors can be sufficient to reorder the leaderboard. To diagnose this dependence, we introduced FST, an information-theoretic estimator that uses marginal and conditional mutual information to identify the direct contributing factor. Building on this diagnosis, FN-VSI reweights instances toward a canonical reference distribution, producing rankings that differ from the standard aggregate and revealing the largest shifts for models whose apparent advantages are most amplified by sensitive factors.

More broadly, our perspective extends beyond VSI-Bench: any benchmark whose instances vary along structured task-agnostic factors raises the question of what its aggregate score actually rewards. Our study focuses on building a diagnostic framework for fairly evaluating existing VLMs, and therefore excludes several competitive models, including Cambrian-S (Yang et al., 2025d) and SenseNova-SI (Cai et al., 2025), whose training data overlap with our evaluation pool. We hope that future scalable data-construction pipelines will enable broader and less contaminated evaluation pools, and that FST can complement these efforts by supporting more interpretable, composition-aware evaluation of spatial reasoning.

Acknowledgements

We thank Jisung Hwang for reviewing the manuscript, and Chanho Park, Jaihoon Kim, and Kyeongmin Yeo for valuable discussions. We also thank the creators of VSI-Bench and VSI-590K for open-sourcing high-quality datasets that made this work possible.

References

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

AI, I., Wang, F., Liu, J., Chen, J., Zhou, J., Ji, K., Ru, L., Guo, Q., Zheng, R., Li, T., et al. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *arXiv preprint arXiv:2507.08306*, 2025.

Anthropic. Introducing claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016.

Asadi, M., O'Sullivan, J. W., Cao, F., Nedaee, T., Fardi, K., Li, F.-F., Adeli, E., and Ashley, E. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS*, 2021.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Brown, E., Ray, A., Krishna, R., Girshick, R., Fergus, R., and Xie, S. Sims-v: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025a.

Brown, E., Yang, J., Yang, S., Fergus, R., and Xie, S. Benchmark designers should" train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025b.

Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

Chen, Y., Zhang, Y., Yu, H., He, J., Wang, Y., Huang, J., Shen, H., Ni, J., Wang, S., Jia, B., et al. Lifting unlabeled internet-level data for 3d scene understanding. *arXiv preprint arXiv:2604.01907*, 2026.

Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., and Huang, R. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

Clark, C., Yatskar, M., and Zettlemoyer, L. 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.

Cohen, J. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *ICCV*, 2025.

Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., and Mottaghi, R. Procthor: Large-scale embodied ai using procedural generation. In *NeurIPS*, 2022.

Deng, N., Gu, L., Ye, S., He, Y., Chen, Z., Li, S., Wang, H., Wei, X., Yang, T., Dou, M., et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025.

Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Feng, H., Lian, L., Dunlap, L., Shu, J., Wang, X., Wang, R., Darrell, T., Suhr, A., and Kanazawa, A. Visually prompted benchmarks are surprisingly fragile. *arXiv preprint arXiv:2512.17875*, 2025.

Gorla, A. and Puduppully, R. The illusion of generalization: Re-examining tabular language model evaluation. *arXiv preprint arXiv:2602.04031*, 2026.

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

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Komanduri, A., Bhaila, K., and Wu, X. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30648–30668, 2025.

Koo, J., Choi, D., Youn, S., Lee, P. Y., and Sung, M. Toward ambulatory vision: Learning visually-grounded active view selection. *arXiv preprint arXiv:2512.13250*, 2025.

Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:62627, 2013.

Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y. R., Lee, S., et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., and Zhuang, Y. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *ICLR*, 2025.

Lian, S., Wu, C., Yang, L. T., Yuan, H., Yu, B., Zhang, L., and Chen, K. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. *arXiv preprint arXiv:2509.24473*, 2025.

Lin, J., Xu, R., Zhu, S., Yang, S., Cao, P., Ran, Y., Hu, M., Zhu, C., Xie, Y., Long, Y., et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025a.

Lin, J., Zhu, C., Xu, R., Mao, X., Liu, X., Wang, T., and Pang, J. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *arXiv preprint arXiv:2507.07984*, 2025b.

Liu, J., Sun, H., Li, W., Zhang, Y., Yang, R., Zhu, Z., Yang, Y., Zheng, S., Jiang, N., Jiang, J., et al. Openspatial: A principled data engine for empowering spatial intelligence. *arXiv preprint arXiv:2604.07296*, 2026.

Ma, X., Sun, T., Chen, S., Bhalgat, Y., Gu, J., Chang, A. X., Armeni, I., Laina, I., Peng, S., and Prisacariu, V. A. Do 3d large language models really understand 3d spatial relationships? *arXiv preprint arXiv:2603.23523*, 2026.

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

Neuhaus, Y., Flammarion, N., Hein, M., and Croce, F. On the out-of-distribution generalization of reasoning in multimodal llms for simple visual planning tasks. *arXiv preprint arXiv:2602.15460*, 2026.

Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.-S., and Wen, J.-R. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., and Sun, X. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., and Ren, Y. C. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *ICCV*, 2023.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Robins, J. 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.

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

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

Singh, S., Nan, Y., Wang, A., D'souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Song, C. H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., and Birchfield, S. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *CVPR*, 2025.

Team, G. R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., 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.

Tong, P., Brown, E., Wu, P., Woo, S., IYER, A. J. V., Akula, S. C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

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

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.

Wu, D., Liu, F., Hung, Y.-H., and Duan, Y. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *NeurIPS*, 2025a.

Wu, H., Huang, X., Chen, Y., Zhang, Y., Wang, Y., and Xie, W. Spatialscore: Towards comprehensive evaluation for spatial intelligence. *arXiv preprint arXiv:2505.17012*, 2025b.

Wu, M., Wang, Z., Wang, F., Yang, J., Pollefeys, M., and Zhang, T. From indoor to open world: Revealing the spatial reasoning gap in mllms. *arXiv preprint arXiv:2512.19683*, 2025c.

Xu, Z., Wang, Z., Wu, J., Lu, J., and Wang, X. Causal-halbench: Uncovering lvlms object hallucinations through causal intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34169–34177, 2026.

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., and Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025a.

Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T. V., Movahedi, M., Li, M., et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *ICML*, 2025b.

Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025c.

Yang, S., Yang, J., Huang, P., Brown II, E. L., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025d.

Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026.

Yang, Y., Lee, C. P., Feng, S., Zhao, D., Wen, B., Liu, A. Z., Tsvetkov, Y., and Howe, B. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? *arXiv preprint arXiv:2506.18322*, 2025e.

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

Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al. Spatial mental modeling from limited views. In *ICLR*, 2026.

Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.-J., Cai, X., Huang, G., et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *NeurIPS*, 2025.

Zhang, J., Zhou, S., Liu, B., Kadambi, A., and Fan, Z. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv preprint arXiv:2603.27437*, 2026a.

Zhang, Y., Li, B., Liu, h., Lee, Y. j., Gui, L., Fu, D., Feng, J., Liu, Z., and Li, C. Llava-next: A strong zero-shot video understanding model, 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Zhang, Y., Chen, J., Tan, J., Mao, Y., Chen, W., and Chang, A. X. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. *arXiv preprint arXiv:2604.24300*, 2026b.

Zhao, C., Tan, Z., Ma, P., Li, D., Jiang, B., Wang, Y., Yang, Y., and Liu, H. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. *arXiv preprint arXiv:2508.01191*, 2025a.

Zhao, R., Zhang, Z., Xu, J., Chang, J., Chen, D., Li, L., Sun, W., and Wei, Z. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025b.

Zheng, D., Huang, S., Li, Y., and Wang, L. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *NeurIPS*, 2025.

A. Appendix

A.1. Details on FN-VSI Reweighting

As described in Sec. 6, we apply two preprocessing steps before computing FN-VSI to improve the reliability and stability of the cell-wise estimates. First, we remove cells with too few instances, since their cell-wise mean scores can be highly unstable while still receiving equal weight in the final factor-normalized average. We use a minimum cell-size threshold of 10 instances. Second, we balance the number of instances across cells to reduce extreme imbalance among the remaining factor configurations. Specifically, we cap overly dense cells by randomly downsampling them with a fixed seed, so that the ratio between the largest and smallest cell counts does not exceed 10.

A.2. Details on Task-Agnostic Factors in VSI-Bench

In this section, we provide further details on each of the task-agnostic factors introduced in Sec. 4. Tab. 2 lists the full set of factors used throughout our analysis.

Input factors. The video data is shaped first by `video_src`—which 3D-scanning dataset the input video belongs to—and derived property: `src_dist`, which captures whether the dataset belongs to VSI-Bench's original source set (ID; ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021)) or lies outside it (OOD; ProcTHOR (Deitke et al., 2022), Hypersim (Roberts et al., 2021), S3DIS (Armeni et al., 2016), ADT (Pan et al., 2023)). On top of this source-level factor, each video carries low-level *visual properties* (`brightness`, `contrast`, and `sharpness`), as well as the video's `file_size`. The depicted `room_type` is also treated as an input-side factor, since it is a property of the captured scene rather than of the spatial reasoning itself.

Query factors. Each question is anchored to an *object of interest* drawn from a fixed vocabulary; the question's effective difficulty depends partly on category-level priors (a *door*, for instance, affords a far stronger size prior than a *box*). We additionally track whether the queried object appears in VSI-Bench's original object set (ID) or outside it (OOD), captured by (`obj_dist`).

Output factors. For tasks with numerical answers, the `gt_val` itself is a property of the benchmark: the distribution of answer values across the evaluation set induces a difficulty axis that is logically independent of the spatial reasoning the question is meant to probe. We also retain a coarsened version, the *ground-truth bin*, to support the analysis that condition on the answer-magnitude regime rather than its exact value.

Binning. Each continuous factor is discretized into task-pooled bins. For `gt_val`, `brightness`, `sharpness`, `contrast`, and `file_size`, we use a tail-aware binning rule: the lower 80 % of values are grouped into one bin, while the top 20 % range is divided into three equal-width bins. This construction always yields four bins and is chosen because these factors exhibit long upper tails, under which equal-frequency quartiles would compress the discriminative high-value range. For `object_type`, we first assign each object category a continuous proxy score based on a held-out proxy model's per-class accuracy on the bounded $[0, 1]$ interval, and then apply equal-frequency quartile cutpoints to preserve balanced sample counts across bins. Borderline cases are assigned using left-closed, right-open intervals, with the maximum value included in the top bin. Because the `object_type` proxy scores contain a heavy tie at one quartile boundary, two adjacent quartile edges collapse after deduplication; therefore, this factor is represented by three bins, whereas all other factors retain four bins. We apply Miller–Madow bias correction using the *observed* cardinality of each factor, which mitigates finite-sample bias and reduces the effect of differing bin counts in cross-factor comparisons.

A.3. Factor Entanglement Analysis

Before we ask how each task-agnostic factor independently affects model accuracy, we first ask whether the factors themselves can be varied independently. If the ten factors of Tab. 2 were uncorrelated, varying one would leave the others unchanged and a per-factor analysis would reduce to ten separate experiments. In practice, however, the factors co-vary in VSI-Bench by construction: scenes from a particular capture pipeline tend to contain particular kinds of objects, particular objects tend to imply particular GT answer ranges. We quantify these pairwise dependencies as our first analytical step.

Measure: pairwise normalized mutual information. For every pair (T_j, T_k) of the ten task-agnostic factors, we report the symmetric arithmetic-normalized mutual information

$$\mathrm{NMI}(T_j, T_k) \;=\; \frac{\widehat{I}(T_j;\, T_k)}{\frac{1}{2}\big[\widehat{H}(T_j) + \widehat{H}(T_k)\big]} \;\in\; [0, 1], \qquad (8)$$

where $\widehat{I}$ is the Miller–Madow bias-corrected plug-in estimator of mutual information and $\widehat{H}$ is the plug-in estimator of Shannon entropy. A value of 0 indicates statistical independence; larger values indicate stronger pairwise dependence. Continuous factors are discretized using the binning rule detailed in Appendix A.2; categorical factors are used at their native level set. We compute the matrix per model and report the mean across the nine VLMs of Tab. 4; per-task effective sample sizes range from N=2,454 (`room size`) to N=19,637 (`object counting`).

Table 2. **Task-agnostic factors in VSI-Bench (Sec. 4).** 10 factors grouped by the aspect of evaluation they govern: input (video data), query (referenced object), and output (answer value). Each factor is a property of the benchmark to which an ideal spatial reasoner should be invariant.

Index	Identifier	Factor	Description
1	video_src	Video source	*Which 3D-scan dataset is the video from?*
2	src_dist	Source distribution	*ID (in VSI-Bench's original sources) or OOD?*
3	brightness	Brightness	*Average pixel intensity over frames*
4	contrast	Contrast	*Image contrast of the frames*
5	sharpness	Sharpness	*Image sharpness over frames*
6	room_type	Room type	*What is the room type?*
7	obj_type	Object type	*Which object is queried?*
8	obj_dist	Object distribution	*ID (object appears in VSI-Bench) or OOD?*
9	gt_val	Ground truth value	*Ground truth answer of the question*
10	file_size	File Size (MB)	*Total file size of the video*

Inter-factor entanglement. Fig. 9 shows the four 9-model-mean NMI matrices, one per task. Three patterns recur consistently. First, the pair video_src–src_dist dominates every matrix (NMI 0.55–0.77); this is by construction, as src_dist is derived from video_src by labelling sources from VSI-Bench's original three datasets (ScanNet, ScanNet++, ARKitScenes) as in-distribution and the rest as out-of-distribution. Second, video_src is moderately to strongly entangled with file_size (NMI 0.34–0.56), reflecting that different 3D-scanning datasets produce videos with systematically different file sizes (encoding parameters, sequence length, resolution). Third, the query-side factors (obj_type, obj_dist) and the answer-side factor gt_val couple to scene-side factors (video_src, src_dist) at task-dependent strengths. The strongest such cross-group pair appears on room size, where gt_val couples to video_src at NMI ≈ 0.34. The remaining pairs—low-level visual statistics (brightness, contrast, sharpness) against query- or answer-side factors—fall below NMI 0.10, indicating that those video-level statistics are largely independent of what is being asked. Tab. 3 lists the top three off-diagonal pairs per task.

A.4. VSI-590K Data Filtering

The fairness of any FST diagnosis hinges on a strict separation between training and evaluation data: an instance used to evaluate a VLM must not have appeared in that VLM's training corpus. Several of the spatial VLMs we evaluate (*e.g.*, SpatialLadder, VLM-3R, Spatial-MLLM) are trained on publicly released spatial QA corpora that, by construction, draw from the same scene datasets as VSI-Bench. Augmenting our evaluation pool with VSI-590K (Yang et al., 2025d), a large-scale training corpus covering the same four numerical tasks, therefore risks contaminating the diagnosis with examples that target models have already seen. We address this with a two-stage overlap filter that removes, from VSI-590K, every instance overlapping with five exclusion datasets at either the surface or the semantic level.

Source corpora. We start from two sources, both normalized to a common schema with video_src, scene_name, question_type, question, gt_value, and a slots field that lists the object(s) referenced by the question: (i) **VSI-590K**, a large-scale training corpus of video-based spatial QA, and (ii) **VSI-Bench** itself, which serves a dual role—both as one of the exclusion datasets used to filter VSI-590K and as a component of the final evaluation pool.

Unit and validity pre-filtering. Per-task pre-filters align units with VSI-Bench and enforce numerical validity. We retain only items whose units match those used in VSI-Bench (object counting: integer counts; object size: centimeters; room size: square meters); object absolute distance: meters, require all ground-truth values to be positive and finite, and apply a 99.5 percentile cap per task to remove extreme outliers that would otherwise distort downstream distribution matching.

Exclusion datasets. We define five exclusion datasets: (i) **SpatialLadder-26K** (Li et al., 2025); (ii) **SR-91K** (Ouyang et al., 2025); (iii) **Spatial-MLLM-120K** (Wu et al., 2025a); (iv) **VLM-3R** (Fan et al., 2025); and (v) **VSI-Bench**, the evaluation benchmark itself. For each exclusion dataset we load all available answer formats (numerical, multiple-choice, open-ended, exact-option) so that overlap is detected regardless of how the same underlying question is phrased. Including VSI-Bench among the exclusion sets guarantees that the filtered VSI-590K subset is, by construction, disjoint from VSI-Bench at the semantic level—this is what allows the final pool to be formed as a simple union without any further deduplication.

Two-stage semantic overlap filtering. Two questions can refer to the same scene–object content even when their surface forms differ (*e.g.*, "*How long is the table?*" vs. "*Estimate the size of the table in cm.*"),

Table 3. Top three off-diagonal factor-factor NMI values per task, averaged over the nine evaluated VLMs. Values are read from the matrices in Fig. 9.

Task	Factor pair	NMI (9-model mean)
object counting	video_src–src_dist	0.66
	src_dist–file_size	0.59
	video_src–file_size	0.56
object size	video_src–src_dist	0.57
	video_src–file_size	0.50
	src_dist–file_size	0.34
room size	video_src–src_dist	0.77
	video_src–file_size	0.34
	src_dist–gt_val	0.32
absolute ddistance	video_src–src_dist	0.55
	video_src–file_size	0.54
	src_dist–file_size	0.36

so a single text-level filter is insufficient. We therefore filter in two stages. *Stage 1 (item-level).* An item is removed if its (scene_source/scene_name, question_text) pair is identical to that of any exclusion item. *Stage 2 (semantic-level).* An item is additionally removed if its *semantic key* matches that of any exclusion item. Semantic keys are built from a normalized scene identifier and the task-specific slots content, as summarized in Tab. 5. Scene identifiers are normalized to absorb variants of the same underlying scene (*e.g.*, scannetpp_v2→scannetpp; scene0191_00→scene0191). Two questions that share scene and target object(s) admit the same ground-truth answer regardless of phrasing or answer format, so a semantic-key match is treated as a duplicate.

Filtering results. Fig. 10 reports per-task filtering outcomes. Each panel pairs a waterfall plot—tracing how each exclusion dataset removes items incrementally when applied in sequence (SpatialLadder-26K, SR-91K, Spatial-MLLM-120K, training dataset from VLM-3R, VSI-Bench)—with a composition chart showing the final retained and removed counts. The retention rate after the full two-stage filter is consistent across tasks: 19,272/24,858 (77.5%) for object counting, 7,993/10,253 (78.0%) for object size, 10,765/15,269 (70.5%) for absolute distance, and 2,166/3,189 (67.9%) for room size, leaving 40,196 filtered VSI-590K instances in total. VLM-3R accounts for the largest share of removals across most tasks (e.g., 11.7%–18.6% on object size and absolute distance), with SR-91K and Spatial-MLLM-120K contributing additional non-trivial overlap especially on object counting and absolute distance, confirming that all four training-side exclusion sources contain content that overlaps semantically with VSI-590K. Notably, vsi_bench removes zero additional items in every task, since the four preceding training-side filters already absorb its semantic content.

Model	Size	Category
Qwen2.5-VL-3B (Bai et al., 2025)	3B	General-purpose
Qwen2.5-VL-7B (Bai et al., 2025)	7B	General-purpose
LLaVA-NeXT-Video-7B	7B	General-purpose
SpatialLadder (Li et al., 2025)	3B	Spatial fine-tuned
Spatial-MLLM (Wu et al., 2025a)	4B	Spatial fine-tuned
SpaceR (Ouyang et al., 2025)	7B	Spatial fine-tuned
VLM-3R (Fan et al., 2025)	7B	Spatial fine-tuned
VST (Yang et al., 2025c)	7B	Spatial fine-tuned
VG-LLM (Zheng et al., 2025)	8B	Spatial fine-tuned

Table 4. Nine VLMs evaluated in this work, grouped by category and ordered by parameter count within each group.

Final pool construction. Let $\widetilde{D}_{\text{590K}}$ denote VSI-590K after unit, validity, and two-stage filtering. The final evaluation pool for each task is

$$\mathcal{P} = \widetilde{D}_{\text{590K}} \cup D_{\text{VSI-Bench}}.$$

Because Stage 2 uses VSI-Bench itself as one of the exclusion sets, $\widetilde{D}_{\text{590K}}$ and $D_{\text{VSI-Bench}}$ are disjoint at both the item and semantic level, so the union requires no additional deduplication. After taking the union with VSI-Bench's 2,640 numerical instances, the final evaluation pool contains 42,836 instances aggregated across the four tasks. All evaluation subsamples used in the main paper are drawn from this pool, ensuring that no item used to evaluate a VLM has appeared in that VLM's training corpus.

A.5. Evaluation Setup Details

Evaluated models. We evaluate nine VLMs spanning general-purpose video VLMs and variants fine-tuned for spatial reasoning, summarized in Tab. 4. Several of the spatial fine-tuned variants share base models with the general-

purpose set (*e.g.*, VST is fine-tuned from Qwen2.5-VL-7B, and Spatial-MLLM from Qwen2.5-VL-3B).

Compute. All experiments are run on a single node with 8×NVIDIA RTX A6000 GPUs.

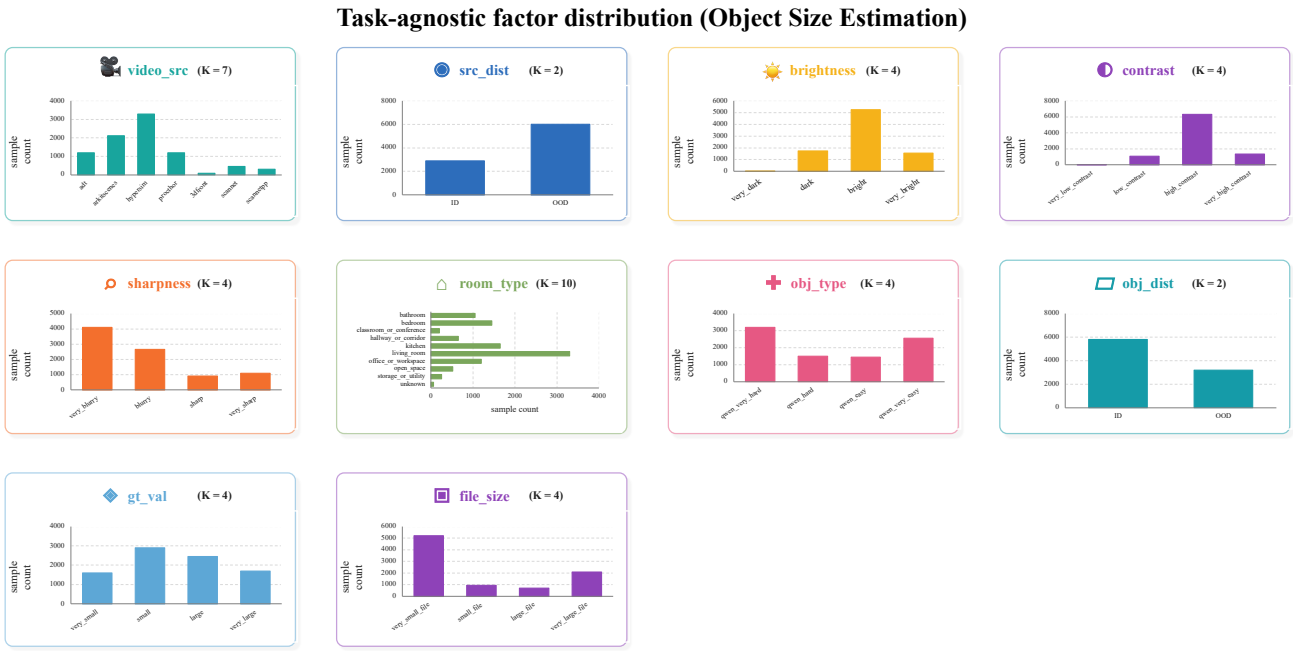

Figure 8. **Task-agnostic factor distribution** on `object size`: per-factor sample counts across bins, with K denoting the number of bins (binning rules in the Binning paragraph below).

Table 5. Task-specific semantic keys used in Stage-2 overlap detection. For `absolute distance`, the two queried objects are sorted alphabetically so that (A, B) and (B, A) collapse to the same key.

Task	Semantic Key	Rationale
`object size`	(scene, target)	single target object
`object counting`	(scene, target)	single target category
`absolute distance`	(scene, sorted(o_1, o_2))	unordered object pair
`room size`	(scene)	scene-level property

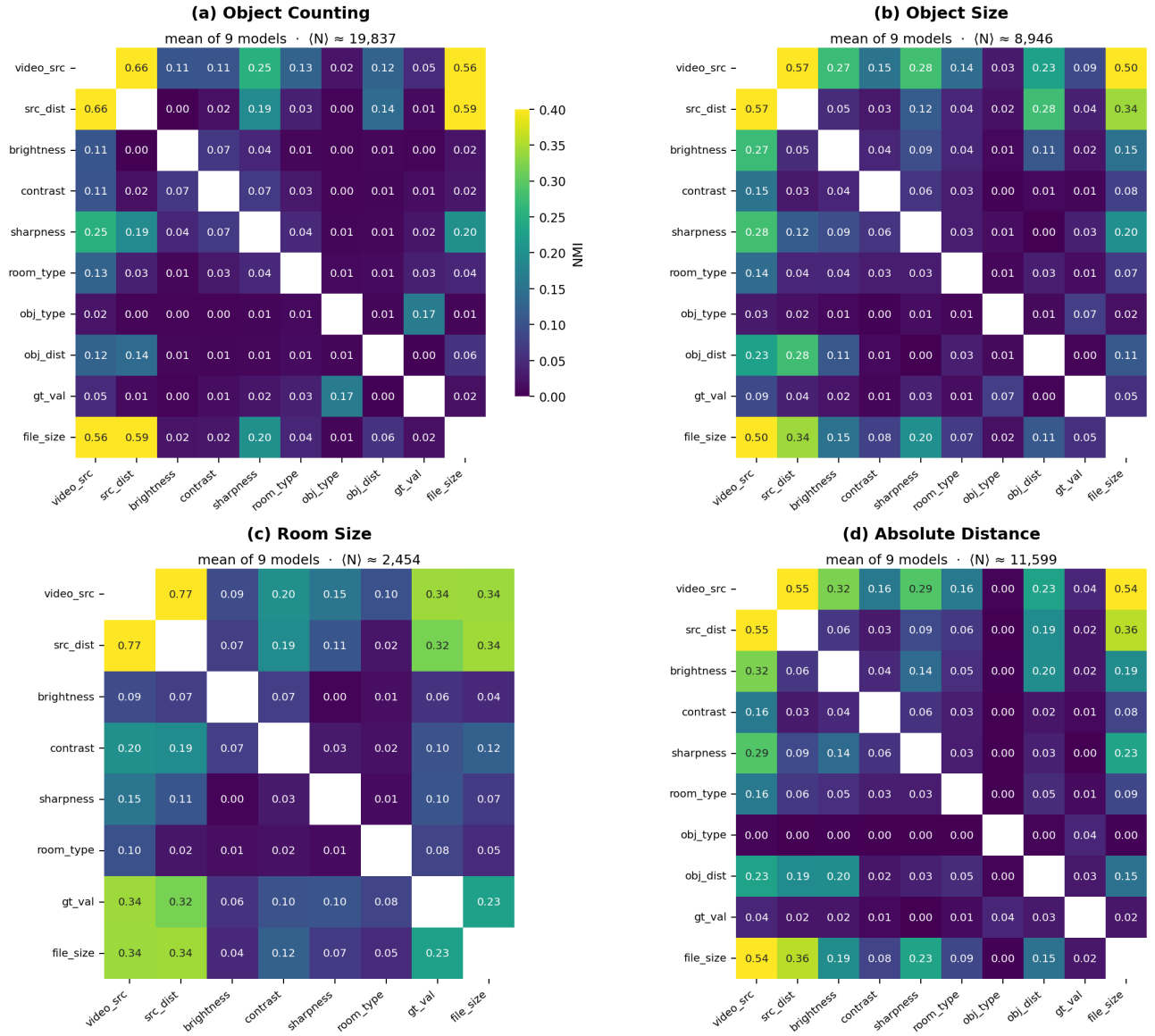

Figure 9. **Inter-factor entanglement on VSI-Bench.** 9-model-mean NMI between the ten task-agnostic factors ((8)), one panel per task. The diagonal is masked; brighter cells indicate stronger dependence.

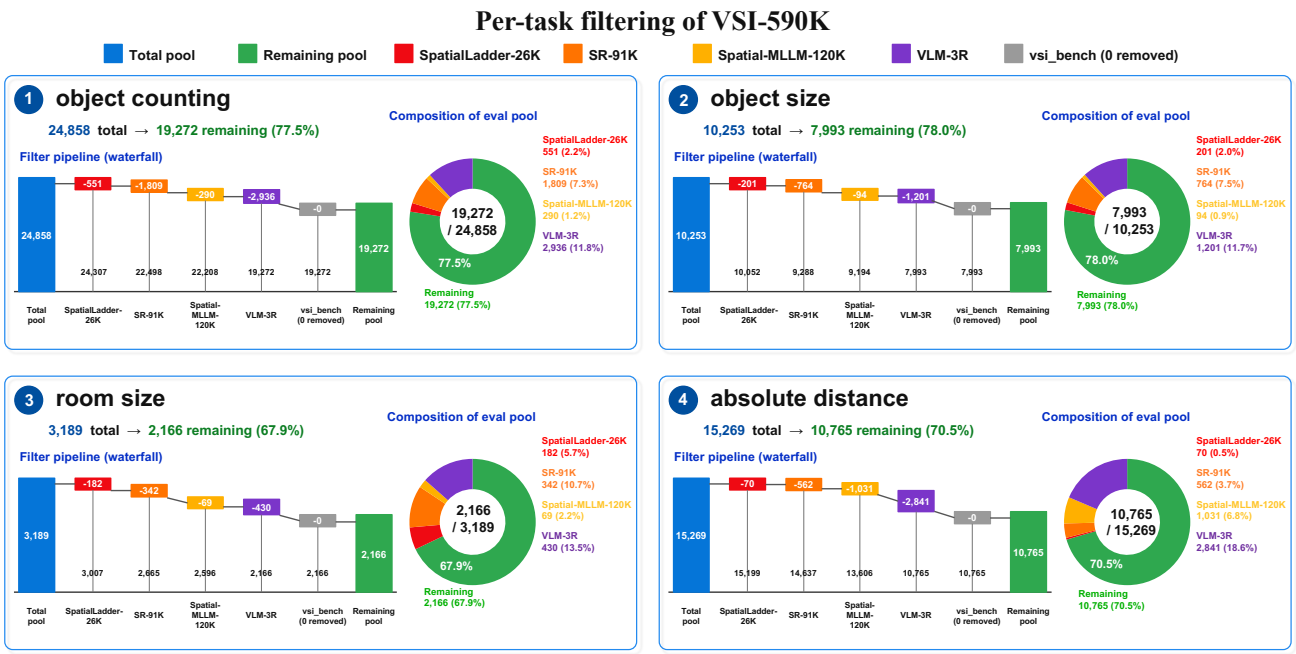

Figure 10. **Filtering VSI-590K to produce $\widetilde{D}_{590K}$.** Each panel shows one task. *Left:* waterfall plot of incremental removals when SpatialLadder-26K, SR-91K, Spatial-MLLM-120K, VLM-3R, and `vsi_bench` are applied in sequence. *Right:* composition chart of the resulting pool, with the contribution of each exclusion source to total removals.

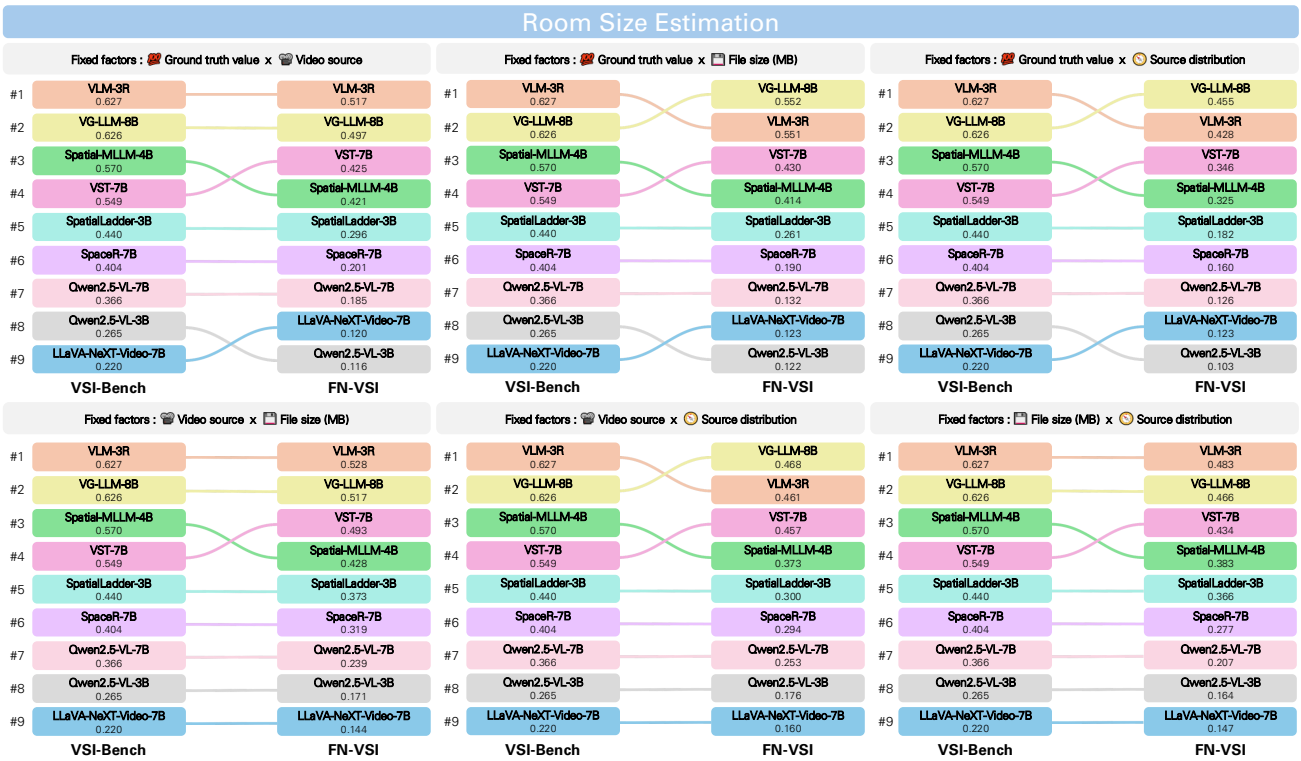

Figure 11. **Ranking shifts under different contributing-factor combinations in Room Size Estimation.** The Room Size Estimation task contains four contributing factors identified by FST: Ground Truth Value, Video Source, File Size, and Scene Distribution. Each panel shows the model ranking obtained after applying FN-VSI using one of the six possible pairwise factor combinations.