Visually Prompted Benchmarks Are Surprisingly Fragile

Anonymous CVPR submission

Paper ID

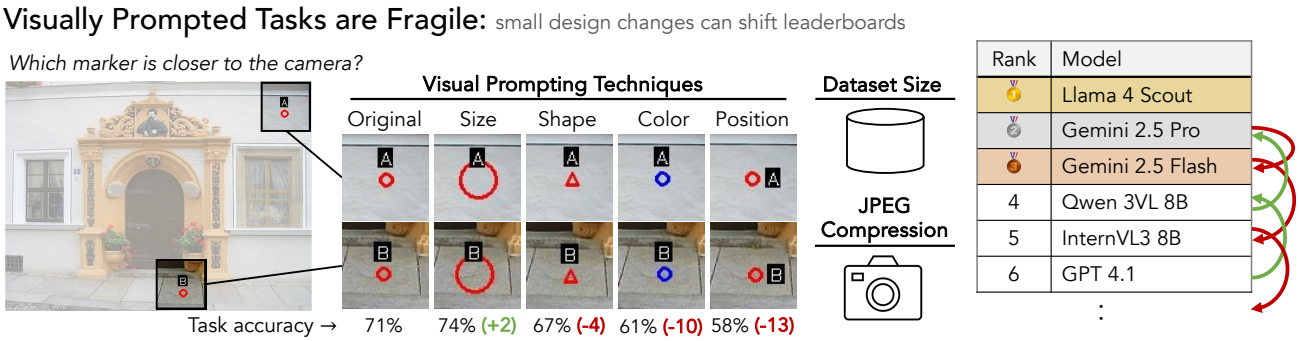


Figure 1. **Small, seemingly irrelevant changes in visual prompting dramatically alter VLM predictions. Left:** Qwen2.5-VL accuracy under different visual marker variants. Changes in marker size, shape, color, and label position lead to accuracy swings of 15%. **Right:** such variations can reorder entire leaderboards, with model rankings shifting even when nothing about the underlying task changes.

Abstract

A key challenge in evaluating VLMs is testing models' ability to analyze visual content independently from their textual priors. Recent benchmarks probe visual perception through visual prompting, where questions about visual content are paired with coordinates to which the question refers, with the coordinates explicitly marked in the image itself. While these benchmarks are an important part of VLM evaluation, we find that existing models are surprisingly fragile to seemingly irrelevant details of visual prompting: simply changing a visual marker from red to blue can completely change rankings among models on a leaderboard. By evaluating nine modern open- and closed-source VLMs on two visually prompted tasks, we demonstrate how details in benchmark setup, including visual marker design and dataset size, have a significant influence on model performance and leaderboard rankings. These effects can even be exploited to lift weaker models above stronger ones; for instance, slightly increasing the size of the visual marker results in InternVL3-8B ranking alongside or better than much larger models like Gemini 2.5 Pro. Furthermore, we find that even ostensibly irrelevant modeling and inference decisions like JPEG compression can change the model lineup while similar interventions to non–visually prompted tasks have little effect on the results. To address this instability, we curate existing datasets to create VPBench, a larger visually prompted benchmark with 16 visual marker variants. We open-source this benchmark as well as our analysis tools to further facilitate robust evaluation of VLMs.

1. Introduction

Despite the rapid progress of vision-language models (VLMs), their visual perception capabilities remain underexplored. Most existing benchmarks conflate visual understanding with language priors and factual recall, making it unclear whether models genuinely perceive or merely retrieve. To address this, visual prompting has emerged as a targeted paradigm: by marking regions in an image and posing spatial or perceptual questions, these tasks assess low-level visual understanding that humans solve effortlessly, in contrast to the knowledge-centric reasoning required by benchmarks such as MME or MMMU [8, 27].

However, within this visually prompted evaluation regime, we find that model performance is surprisingly sensitive to seemingly minor design choices in the bench-

mark itself. As illustrated in Figure 1, variations in the size, style, or layout of visual markers can substantially affect accuracy and even reorder model rankings. Beyond prompt design, incidental implementation details, such as random sample selection, image compression settings, or floating-point precision, can further contribute to this instability. Many of these design choices are inherited from conventional, knowledge-focused VLM benchmarks, where such factors have minimal influence and are thus treated as inconsequential. Yet, in visually prompted evaluations, these non-semantic elements become hidden confounders, capable of markedly distorting model performance and leaderboard rankings. Consequently, existing visually prompted benchmarks exhibit an inherent fragility—blink again, and an apparently incidental change can shift reported scores, echoing the formatting sensitivities observed in LLMs [18]. Such instability undermines confidence in benchmark-driven progress and echoes recent concerns about leaderboard fragility in both language and vision domains.

We explore three such sources of evaluation instability across eight modern VLMs on three visually prompted datasets (BLINK [9], DA2k [25], and SPair [15]). First, we demonstrate the effect of *sample choice*: random resampling of image subsets, matched in size to BLINK and drawn from a fixed pool of visually prompted tasks, leads to substantially reordered model rankings, despite the subsets being statistically indistinguishable in size and difficulty. Second, we examine *visual prompt formatting*. For BLINK, DA2k, and Spair, we evaluate our models with 16 different marker styles varying in size, shape, color, and label placement. We find that marker style can cause accuracy swings of up to 21% on the same image-question pairs, often causing ranking reversals among state-of-the-art models. Finally, we show that that *inference-time implementation details* which are imperceptible to humans such as JPEG compression further perturb results in statistically significant ways. Additionally, we find that this is specific to visually prompted tasks, as applying the same intervention to more traditional VLM benchmarks does not significantly change the results. We also demonstrate how this fragility can be exploited to "game" leaderboards. For example, strategically selecting the visual marker to be a square instead of a circle causes a weaker model like InternVL3-8B to rank above stronger models like Gemini 2.5 Pro on the BLINK relative depth estimation task.

These findings suggest that much of the variation among model performance reported on vision-language benchmarks comes not from differentiated intrinsic capabilities of grounding language in vision, but from incidental details of prompting, sampling, and implementation. To address this, we release larger versions of the BLINK, DA2k, and SPair dataset, covering 16 different visual marker variants, and overall boosting the dataset size from 224 samples in BLINK relative depth and semantic correspondence to 35,088 annotated images across both tasks. Lastly, we provide suggestions on how to more robustly evaluate visually prompted tasks and guidance on when to trust the leaderboard rankings. We release code, data, and evaluation scripts to facilitate adoption and to help the community build more robust benchmarks.

To facilitate future research, we will release our proposed VPBench along with our inference code, which supports varying visual markers and image compression settings, as a reference inference for stable evaluation.

2. Related Work

Perception-focused VLM benchmarks. Our work builds on efforts to separate low-level VLM perception from high-level reasoning. BLINK, for example, recasts classic vision tasks into visually prompted questions, showing that VLMs struggle on problems humans solve "in a blink" [9], motivating follow-up work on perception-augmented representations [3]. Orthogonal work evaluates robustness via controlled input variations, such as perturbations to image–text pairs, programmatic generation of task variants, geometric invariances, or spatio-temporal video manipulations [1, 7, 16, 30]. These benchmarks highlight VLM gaps in both perception and robust invariance.

Text and *visual* prompt sensitivity. Prompt design can dominate model performance. In language models, small, meaning-preserving formatting tweaks can significantly swing accuracy [18]. This sensitivity extends to the *visual* prompts in VLMs, where the choice of marker (e.g., a dot vs. a box) can alter model attention and outcomes [19]. Indeed, the space of visual prompts has itself become an optimization target to boost accuracy [29]. These findings imply that evaluations using a single visual style may reflect prompt idiosyncrasies more than true model competence.

Evaluation instability: leaderboard fragility and implementation subtleties. Evaluation outcomes for VLMs are highly unstable, sensitive both to benchmark design and to low-level implementation details. *On the benchmark side*, seemingly minor factors such as altering multiple-choice option order or varying random seeds can flip leaderboard rankings [2, 14]. Community leaderboards can also be gamed via selective submissions and feedback loops [20], motivating automated pipelines that continuously refresh and diversify test sets [13]. *On the implementation side*, minor choices can likewise skew results: in image generation, resizing filters or JPEG compression significantly impact FID scores [17], while in multimodal evaluation, imperceptible perturbations or numerical precision differences

(FP16 vs. FP32) can destabilize outputs and compromise reproducibility [10, 23, 26]. Together, these findings underscore that leaderboard orderings may reflect artifacts of evaluation pipelines as much as genuine model ability.

Spatial reasoning benchmarks and methods. Spatial reasoning has been a long-standing challenge, from early synthetic benchmarks [11, 22] to modern evaluations. Recent studies consistently show that state-of-the-art VLMs fail on simple spatial tasks, often performing near chance on benchmarks probing relative positioning and grounding [12, 21, 24]. To close this gap, a new wave of work introduces spatially-aware architectures and large-scale, spatially-grounded training data [4–6].

In summary, while prior work has documented fragility in *textual* prompting and robustness under broad visual corruptions, we focus on the *fine-grained, visually prompted* regime—benchmarks that explicitly mark regions or points in an image to elicit low-level perceptual judgments (e.g., relative depth). We demonstrate that (i) visual prompt style is a primary confounder, (ii) i.i.d. resampling from a fixed superset can reorder model rankings , and (iii) low-level implementation choices exacerbate variance. To counteract this, we advocate confidence-aware reporting across diversified visual prompts and stratified resamplings—analogous in spirit to FormatSpread's multi-format evaluation, but tailored to VLM perception—which yields markedly more stable rankings on BLINK-like tasks.

3. Visually Prompted Tasks and Probing Them

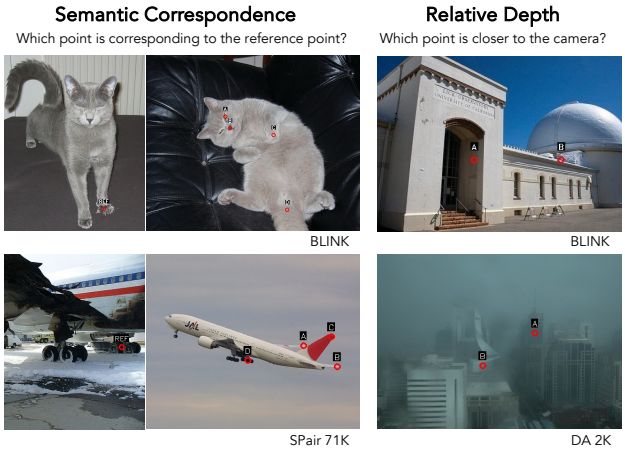


Figure 2. **Examples of visually prompted tasks.** Visually prompted tasks (VPTs) involve placing visual markers in the image to ask questions such as relative depth and semantic correspondence.

3.1. What are Visually Prompted Tasks?

A *visually prompted task* explicitly marks regions of an image and asks about relationships within or between images [9]. We picked two typical visually prompted tasks (Fig. 2). In *semantic correspondence*, two images with marked points are shown and the model is asked which marker corresponds to the same object part or region as the reference marker. In *relative depth*, two locations in an image are marked with "A" and "B," and the model is asked which is closer to the camera.

Functionally, such visual prompting complements verbal prompting, serves as an intuitive and effective interface for querying or referencing fine-grained visual content. Humans naturally point to a location on a map, for example, rather than verbally describing its longitude and latitude. Meanwhile, unlike most VLM tasks that depend on broad world knowledge (e.g., MMMU), these visually prompted tasks are perceptual: humans solve them "in the blink of an eye," relying on fine-grained visual reasoning rather than factual recall. This makes them a natural choice for evaluating how well models recognize objects and spatial relationships without external knowledge. BLINK [9], the benchmark we study, is built around such tasks and has become a de-facto standard for measuring visual perception in VLMs.

3.2. Experimental Setup

The BLINK dataset [9] systematizes these visually prompted tasks via curated image–question pairs with region-level markings. It contains relative depth and correspondence tasks with high annotation quality. However, we noticed that within BLINK, visual markers are not standardized: some examples use dots, others boxes or arrows. When these seemingly cosmetic differences are altered, we noticed non-trivial change in model performance. This motivated our investigation into how VLM benchmarks may be fragile given seemingly irrelevant design choices.

As will be detailed in Section 4, BLINK's small size, approximately 100 examples per split, makes it difficult to separate true performance variation from sampling noise. To reduce this ambiguity and enable more robust analysis, we construct two larger datasets that follow the same visually prompted task format. DA2K, originally introduced for depth annotation, contains thousands of images with dense geometric labels; we repurpose it into a relative-depth task by converting pixel-level depth into pairwise comparisons following BLINK's protocol. SPair-71k (SPair), designed for semantic correspondence, provides tens of thousands of image pairs with detailed keypoint matches; by adopting BLINK's prompting style, we frame SPair as a visually prompted correspondence benchmark that bridges traditional CV evaluation with modern VLM usage. We then apply all interventions for the following sections to all 4 datasets.

Additionally, we compare the instability of rank across different dataset sizes and JPEG compressions to a non visually prompted task, MME [8], to investigate how these interventions affect visually prompted tasks specifically.

We evaluate four closed-source and five open-source VLMs: Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4.1, GPT-4o, Llama 4 Scout, Qwen3-VL-8B, Qwen2.5-VL-7B, Gemma 3-4B, and InternVL3-8B. Aggregate accuracies for each benchmark appear in Fig. 3.

In the following section, we demonstrate the fragility of these visually prompted tasks by showing how small changes in data sampling, visual marker, and low-level implementation details can completely change the results .

4. Statistically Equivalent Sampling Yields Different Results

When constructing a benchmark, creators typically sample a subset of data points from a large data pool, such as data collected from the Internet, to form the evaluation set. This sampling process is random, so in principle, any subset drawn from the same pool should be statistically equivalent and yield similar results. The subset size is usually fixed by convention, following prior knowledge-oriented VLM benchmarks where per-task sample counts commonly range from 50 to 500 items. While BLINK also follows this convention, we find that the random sampling step itself can meaningfully influence evaluation outcomes. If a different subset were drawn from the same underlying distribution, both the absolute accuracy and the model ranking can change noticeably.

Experimental setup: We create 1,000 new BLINK size datasets by randomly sampling 100 samples from each of DA2k [14], Spair 71k [15], and the non-visually prompted dataset MME [8]. Since these samples are drawn from the same underlying data distribution, the accuracies and model leaderboard should remain constant, but as shown in Figure 4, we can get a complete change in rankings across DA2k and Spair.

Why sampling matters here: This sensitivity arises because BLINK, like many newly-released multimodal benchmarks, deliberately targets *unsaturated* capabilities performance levels far from ceiling, typically in the 30–60% range rather than the 80–90% common in relative mature VLM tasks, consequently, these VPT accuracy deviation would be much higher than knowledge-focused ones like MME (See Fig. 5). At such scales , even a few items can noticeably shift results: in BLINK, a 3-point accuracy change corresponds to only two or three additional correct responses, which in a multiple-choice format could easily arise by chance. The resulting variance is reflected in the confidence intervals in Figure 3, and we further highlight its impact by showing how results fluctuate as the dataset size decreases.

Size particularly matters for VPTs: Additionally, we compute the mean standard deviation in model performance across splits to get numeric measures of the instability seen across splits. Table 5 shows that the instability seen across these splits is considerably larger than that of non visually prompted tasks. This suggests that more so than for other visual tasks, visually prompted benchmarks would actually require a relatively larger number of samples to reduce the variance you see from the data. In the following sections, we will show how additional choices effect the leaderboard even when given a larger dataset size.

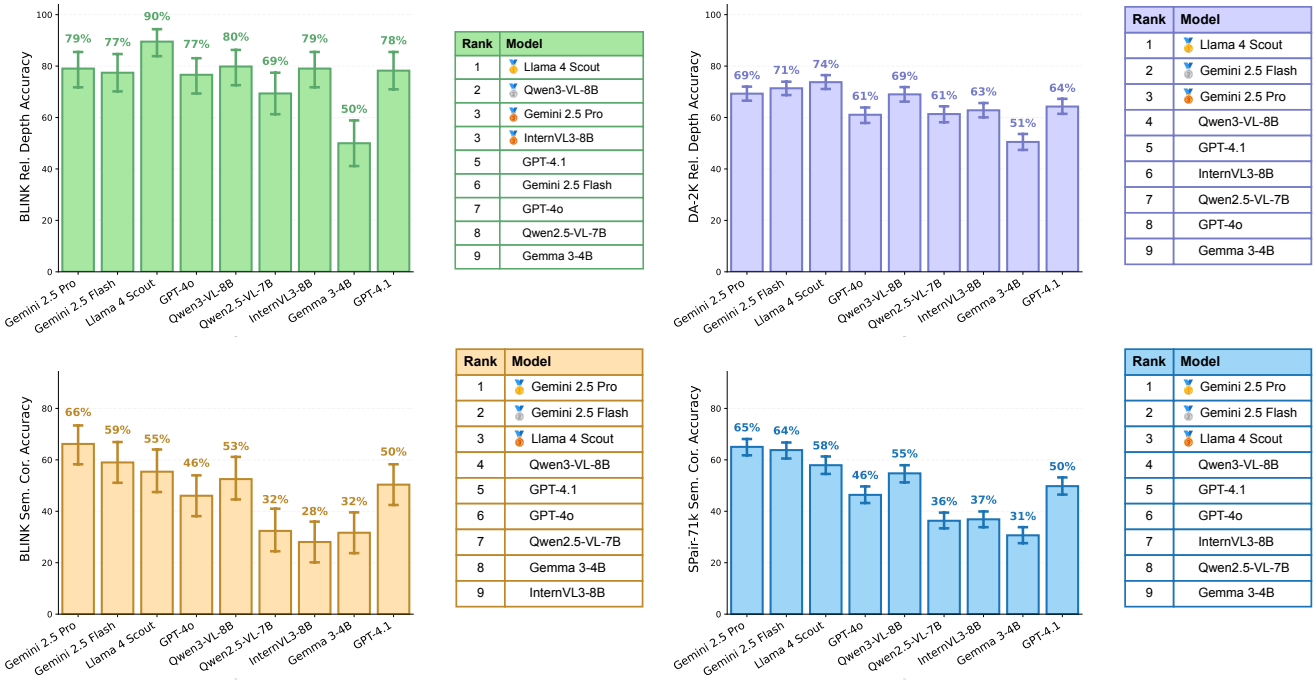
5. Visual Marker Styles Shuffle Leaderboards

We investigate whether the style of the visual marker used in a prompt influences model performance, and whether it can change the relative ordering of models' accuracy. This question is motivated by an observation of inconsistent marker styles in the BLINK benchmark dataset [9]. The visual prompts BLINK employs as part of the question context occasionally switched in color or text positioning and we noticed, for example, that simply switching a marker's color from red to blue questions led to measurable changes in a model's accuracy on BLINK, DA2k, and SPair, suggesting current models are over-reliant on specific visual cues. In this section we create a set of marker variants and study their effects on model performance. We report results on our larger datasets of SPair and DA2k in the main paper and report results for the BLINK subsets in the Appendix.

Experimental setup: Based on the most common marker seen in BLINK relative depth and semantic correspondence, we define a default marker style to be a small red circular marker with a numeric label placed above. We then create 16 alternative styles spanning 4 categories: color (blue instead of red), shape (square instead of circle), size (larger marker, radius increased from the default to size 10), and text position (label moved to below the marker). Figures 6a and 6b display the change in accuracies and rankings of each model on DA2k and SPair datasets.

Disentangling visual marker variance from data variance. To confirm that the variance in accuracy seen across visual markers is not simply due to the variance in the data itself, we perform a paired bootstrap to confirm that the difference between the default BLINK marker and other marker variance is statistically significant.

Let $\mathcal{D} = \{(x_\ell, y_\ell)\}_{\ell=1}^{N}$ be the dataset. A marker m renders an annotated input $m(x_\ell)$. BLINK provides a default



Rank	Model
1	Llama 4 Scout
2	Qwen3-VL-8B
3	Gemini 2.5 Pro
3	InternVL3-8B
5	GPT-4.1
6	Gemini 2.5 Flash
7	GPT-4o
8	Qwen2.5-VL-7B
9	Gemma 3-4B

Rank	Model
1	Llama 4 Scout
2	Gemini 2.5 Flash
3	Gemini 2.5 Pro
4	Qwen3-VL-8B
5	GPT-4.1
6	InternVL3-8B
7	Qwen2.5-VL-7B
8	GPT-4o
9	Gemma 3-4B

Rank	Model
1	Gemini 2.5 Pro
2	Gemini 2.5 Flash
3	Llama 4 Scout
4	Qwen3-VL-8B
5	GPT-4.1
6	GPT-4o
7	Qwen2.5-VL-7B
8	Gemma 3-4B
9	InternVL3-8B

Rank	Model
1	Gemini 2.5 Pro
2	Gemini 2.5 Flash
3	Llama 4 Scout
4	Qwen3-VL-8B
5	GPT-4.1
6	GPT-4o
7	InternVL3-8B
8	Qwen2.5-VL-7B
9	Gemma 3-4B

Figure 3. **Model performance across BLINK, DA2k, and SPair.** Accuracy and rankings of 9 VLMs on BLINK Relative Depth, BLINK Semantic Correspondence, DA2k Relative Depth, and SPair-71k Semantic Correspondence using BLINK's default marker conventions. Error bars are 95% confidence intervals.

marker m_0; we also consider valid alternatives $m_1, \ldots, m_K$ (e.g., dots, boxes, arrows). For a model f, accuracy under marker m is

$$\mathrm{acc}_m = \frac{1}{N} \sum_{\ell=1}^{N} \mathbf{1}\big[f(m(x_\ell)) = y_\ell\big].$$

We draw B bootstrap replicates $\mathcal{D}^s$ by sampling N items with replacement. On each replicate, we compute accuracies with the default and an alternative marker and take the paired difference

$$\Delta\mathrm{acc}_j^s = \mathrm{acc}_{m_0}^s - \mathrm{acc}_{m_j}^s.$$

The distribution $\{\Delta\mathrm{acc}_j^s\}_{s=1}^B$ isolates marker variance because both conditions use the same sampled items. We test $H_0 : \mathrm{acc}_{m_0} = \mathrm{acc}_{m_j}$ by checking whether zero lies in the 95% bootstrap confidence interval of $\Delta\mathrm{acc}_j^s$.

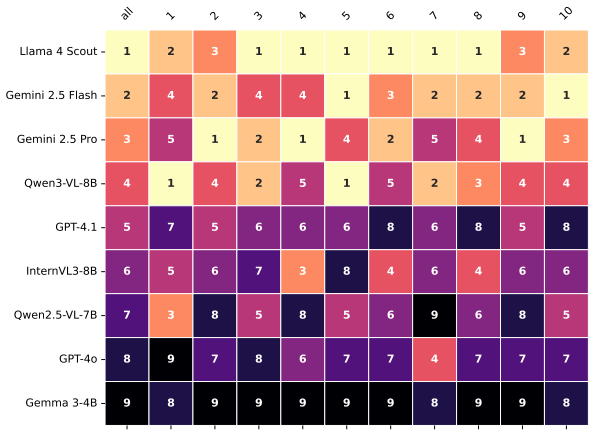
To compare sources of variability, we report the ratio

$$R = \frac{\mathbb{E}_j\big[\mathrm{Var}\big(\Delta\mathrm{acc}_j^s\big)\big]}{\mathrm{Var}\big(\mathrm{acc}_{m_0}^s\big)},$$

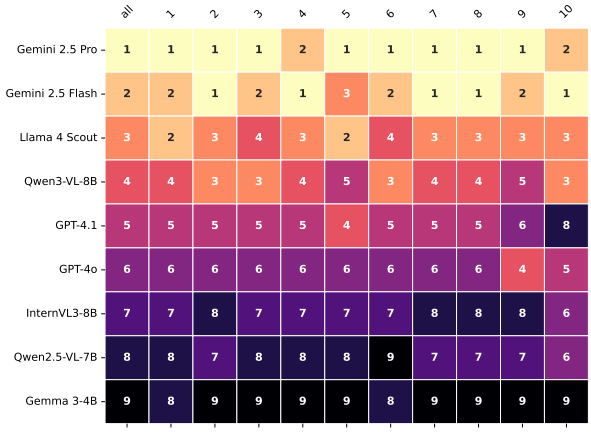
where the denominator estimates dataset variance and the numerator estimates marker-induced variance. Values $R \geq 1$ indicate that changing marker style perturbs accuracy at least as much as resampling the test set itself. We find the majority of visual markers show significant differences for at least 1 model, proving that marker style is not cosmetic but consequential (more details in the Appendix).

Effect on model rankings: Figures 6a and Figure 6b show the model accuracy on the default visual marker (red circle, text at the top) along with the delta in performance seen across 16 different visual markers of varying sizes, shapes, colors, and positions. We see that if you change the visual marker, the takeaways about which model is superior can be vastly different. Beyond absolute accuracy, marker changes often altered the relative performance of models. For example, Llama 4 Scout outperformed Gemini 2.5 Flash under the default red marker for DA2k, but under the blue marker condition their ranks reversed, with Gemini 2.5 Flash scoring as the highest performing model. We also see cases of incredibly large differences in accuracy, such as Llama 4 Scout and InternVL in SPair, where making the text size of the marker smaller results in a 10+% drop in accuracy and dropping Llama from rank 3 to rank 6.

Which visual markers matters most? Figure 7 displays the average magnitude of accuracy difference across each marker for the DA2k dataset. altering size or label position generally produced larger effects on accuracy than changing



(a) DA2k Model Ranking per data split



(b) Spair Model Ranking per data split

Figure 4. **Model rankings across 1,000 independent 100-sample splits for DA2k and SPair-71k**, with the first 10 splits visualized, revealing substantial ranking volatility caused solely by i.i.d. resampling.

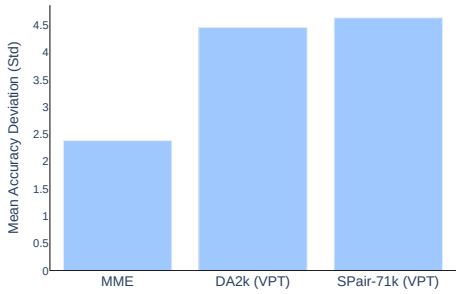


Figure 5. **Model accuracy change across 1,000 splits of 100 samples.** Standard deviation in accuracy averaged across models. We see higher rank instability across 1,000 subsets each with 100 samples for the visually prompted tasks (DA2k and SPair) compared to knowledge based VLM tasks (MME).

color or shape. Changing the marker's color (red to blue) had a large impact on certain models, hinting that those models might be overfit to the color distribution of markers seen in their training or the benchmark (since BLINK's default is red, some models may have learned to specifically attend to red circles as a prompt cue). Changing the marker's size and shape or changing the text marker from A/B to 1/2 had a measurable but less prominent effect on accuracy. Notably, we did not find a single marker style that was universally best or worst for all models. These idiosyncratic responses point to each model having its own biases in visual prompt processing.

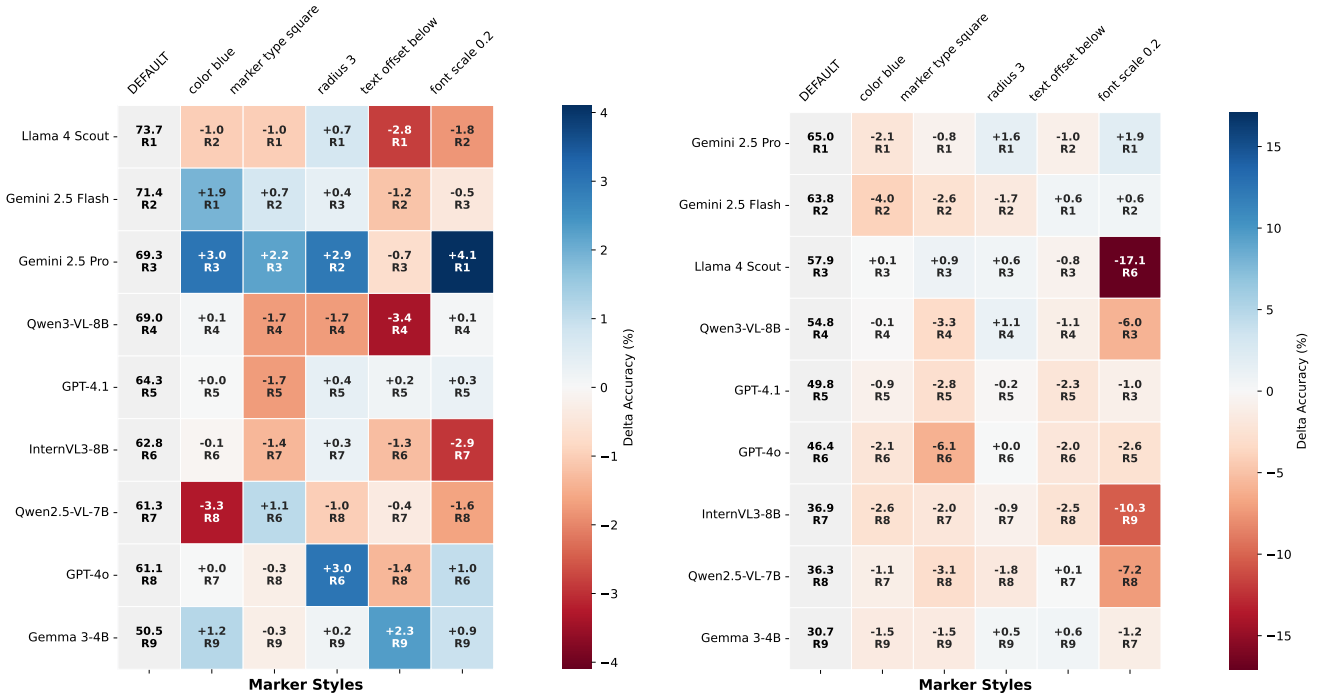
Manipulating Visually Prompted Leaderboards. We explicitly demonstrate that visually prompted leaderboards can be *gamed* by jointly selecting a marker style and an i.i.d. split that favor a particular model. On BLINK relative depth, for example, we can lower internVL3 from rank 4 to rank 8 by simply changing the marker to be a square, and can similarly raise its rank to 3, higher than Gemini 2.5 Pro, by increasing the font size of the marker. This underscores that, without standardized marker conventions, visually prompted evaluations can give misleading impressions of model ability.

6. Imperceptible difference matter as well

Beyond visually noticeable details such as marker styles, we also extend our investigation to *imperceptible* factors, drawing inspiration from prior work on human perceptual sensitivity and adversarial robustness [*e.g.*, 28]. In particular, we examine whether common preprocessing operations—such as JPEG compression, which is widely applied in benchmark construction for efficient storage—can subtly affect model performance. Although variations in JPEG quality above a compression level of 70 are largely imperceptible to humans, it remains unclear whether vision–language models (VLMs) exhibit comparable robustness. Moreover, as this issue has not been systematically explored in prior literature, we investigate whether such imperceptible variations affect vision–perception tasks (VPT) differently from conventional knowledge-focused benchmarks.

Setup. We evaluate four different *JPEG compression levels*: default, 70, 80, and 90. Prompts and data splits are kept fixed. To emphasize rank stability rather than absolute performance, Fig. 6 reports *model ranks* (1,=,best) instead of raw accuracies—(a) for BLINK–Relative Depth (RD) and (b) for MME (semantic).

BLINK RD (VPT) vs. MME (semantic). Surprisingly—or perhaps not surprisingly—the rankings in BLINK RD fluctuate considerably, even for closed-source models such as Gemini 2.5 Pro. This sensitivity can be attributed to the fact that VPT tasks demand a more fine-grained understanding of visual tokens than semantic-focused tasks;



(a) **Change in model accuracy for different visual markers on DA2k.** (b) **Change in accuracy for different markers on SPair.**

Figure 6. Accuracy deltas for DA2k (a) and SPair-71k (b) across 16 visual marker variants, with five selected for visualization, demonstrating strong sensitivity to small changes in marker appearance.

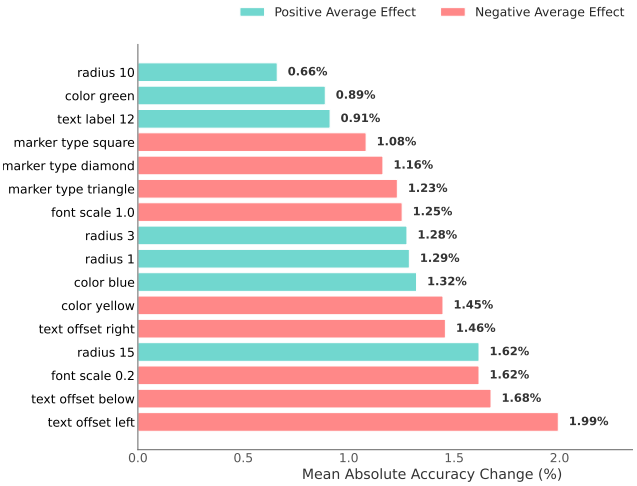


Figure 7. Mean absolute accuracy shift on DA2k induced by each marker variant.

consequently, even subtle pixel-level changes can influence model predictions. In contrast, MME rankings remain remarkably stable across all four compression settings and nine evaluated models, with only a single ranking inversion compared to the pronounced shifts observed in BLINK RD (see Fig. 9b for details).

Takeaways. JPEG quality should be standardized—or ideally replaced with lossless formats—for VPT benchmarks; otherwise, compression artifacts alone can meaningfully tilt leaderboard standings when model performances are tightly clustered.

7. Discussions

Our results show that visually prompted evaluations introduce non-semantic confounders: marker design, sample choice, and low-level implementation details can all shift accuracies and reorder rankings. This fragility arises both from the benchmarks and from the models themselves.

To partially mitigate these issues, we advocate the following practices:

- **Standardize and diversify visual prompts.** Use a clearly specified default marker style, and, whenever possible, report results averaged over a small set of marker variants. For benchmark creators, release clean images together with raw marker coordinates rather than only images with markers baked in, so that alternative prompt designs can be evaluated consistently.
- **Enrich test sets from consistent sources.** When feasible, evaluate on larger visually prompted pools constructed from the same underlying data sources and task definitions, rather than relying on a single small split. In

Default
Standard Evaluation

Model	Rank	Score
Llama 4 Scout	#1	89.41%
Gemini 2.5 Flash	#2	77.65%
Gemini 2.5 Pro	#2	77.65%
InternVL3-8B	#4	76.47%
Qwen3-VL-8B	#4	76.47%
GPT-4.1	#6	75.29%
GPT-4o	#7	71.76%
Qwen2.5-VL-7B	#8	70.59%
Gemma 3-4B	#9	52.94%

Deflate InternVL3-8B
Marker Type Square

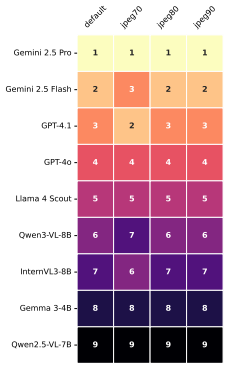
Model	Rank	Score
Llama 4 Scout	#1	90.59%
Gemini 2.5 Pro	#2	84.71%
Qwen3-VL-8B	#3	77.65%
GPT-4.1	#3	77.65%
Gemini 2.5 Flash	#5	76.47%
GPT-4o	#6	75.29%
Qwen2.5-VL-7B	#7	67.06%
InternVL3-8B	#8	63.53%
Gemma 3-4B	#9	55.29%

Inflate InternVL3-8B
Font Scale 1.0

Model	Rank	Score
Llama 4 Scout	#1	85.88%
Gemini 2.5 Flash	#2	81.18%
InternVL3-8B	#3	77.65%
Qwen3-VL-8B	#3	77.65%
Gemini 2.5 Pro	#5	76.47%
GPT-4o	#6	72.94%
GPT-4.1	#7	71.76%
Qwen2.5-VL-7B	#8	70.59%
Gemma 3-4B	#9	49.41%

Figure 8. Performance comparison: optimizing for InternVL3-8B's ranking on BLINK relative depth.

(a) **BLINK RD:** change in accuracy across JPEG qualities.



(b) **MME:** ranks are largely invariant to JPEG compression rates.

Figure 9. Effect of JPEG compression.

our case, we expand BLINK-style relative depth and correspondence tasks using DA2k and SPair-71k, yielding substantially more stable aggregates.

- **Adhere to the same realization of low-level settings.** Explicitly standardize and report data-processing and inference choices such as JPEG compression quality, input resolution, and numerical precision (e.g., `bf16` vs. `fp8`/`fp16` for self-hosted models). Avoid silent changes across evaluations, and treat models with opaque internals as a separate comparison group.
- **Report uncertainty and rank stability.** Accompany accuracies with confidence intervals (e.g., Wilson or bootstrap) and simple rank-stability analyses across markers, splits, or seeds. When intervals overlap substantially, treat models as effectively tied rather than declaring definitive winners.

As a step toward more stable evaluation, we will release VPBench, the scaled-up visually prompted benchmark used in our work, together with multiple marker variants and reference evaluation scripts. By reducing variance due to arbitrary design choices, VPBenchmakes performance differences more reflective of true perceptual ability. While currently limited to depth and correspondence tasks, it provides a foundation for broader analyses and future robustness-aware benchmarks for visual grounding.

8. Conclusion

Benchmarks should measure ability, not fragility. Yet our results show that visually prompted evaluations fail this : change a marker's color, shift its label, compress an image differently, and entire leaderboards reshuffle. These shifts are not noise but structural weaknesses, revealing that today's perception-focused VLM benchmarks are far more sensitive than the field assumes.

Although our demonstrations center on BLINK-style tasks, the pattern is unlikely to be isolated. Any benchmark that depends on explicit visual markup or fine-grained spatial cues risks similar instability. If leaderboards can be flipped by choices orthogonal to task semantics, they cannot be trusted to track genuine progress.

We argue that evaluation must evolve accordingly: diversify visual prompts, report variance in addition to scores, and standardize low-level settings that silently influence results. VPBench is a step in that direction, offering larger, marker-diverse test sets that reduce incidental variance. Stable measurement is a prerequisite for meaningful comparison; until then, visually prompted leaderboards may be telling us more about their construction than about the models they rank.

References

[1] Amit Agarwal, Srikant Panda, Angeline Charles, Bhargava Kumar, Hitesh Patel, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, et al. Mvtamperbench: Evaluating ro-

bustness of vision-language models. *arXiv preprint arXiv:2412.19794*, 2024. 2

[2] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, 2024. 2

[3] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025. 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024. 3

[5] 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. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

[6] Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025. 3

[7] Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R Fung. Unveiling the lack of lvlm robustness to fundamental visual variations: Why and path forward. *arXiv preprint arXiv:2504.16727*, 2025. 2

[8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 4

[9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2, 3, 4

[10] Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/. 3

[11] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017. 3

[12] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 3

[13] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024. 2

[14] Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024. 2, 4

[15] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 2, 4

[16] Seulki Park, Daeho Um, Hajung Yoon, Sanghyuk Chun, and Sangdoo Yun. Rococo: Robustness benchmark of ms-coco to stress-test image-text matching models. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024. 2

[17] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 2

[18] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023. 2

[19] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023. 2

[20] Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025. 2

[21] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025. 3

[22] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 3

[23] Jordan Vice, Naveed Akhtar, Yansong Gao, Richard Hartley, and Ajmal Mian. On the reliability of vision-language models under adversarial frequency-domain perturbations. *arXiv preprint arXiv:2507.22398*, 2025. 3

[24] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024. 3

[25] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2

[26] Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Give me fp32 or give me death? challenges and solutions for reproducible reasoning. *arXiv preprint arXiv:2506.09501*, 2025. 3

[27] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. 1

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 6

[29] Yuan Zhang, Chun-Kai Fan, Tao Huang, Ming Lu, Sicheng Yu, Junwen Pan, Kuan Cheng, Qi She, and Shanghang Zhang. Autov: Learning to retrieve visual prompt for large vision-language models. *arXiv preprint arXiv:2506.16112*, 2025. 2

[30] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 2

Supplementary Materials

A. Marker style significance

Figure 10 reports which marker-induced differences on SPair-71k and DA2k are statistically significant under paired confidence intervals, we can see that all the models have at least one style that is statistically significant, justify the independence of the marker style variance over the data variance. We additionally show the mean accuracy shift per marker on SPair in Figure 11, in which we see that similar to the results from DA2k in the main paper (Figure 7), the marker variants with the largest effects are the ones which involve changing the placement, size, and representation of the text of the marker. In Figure 12, we illustrates the marker variants used in our experiments.

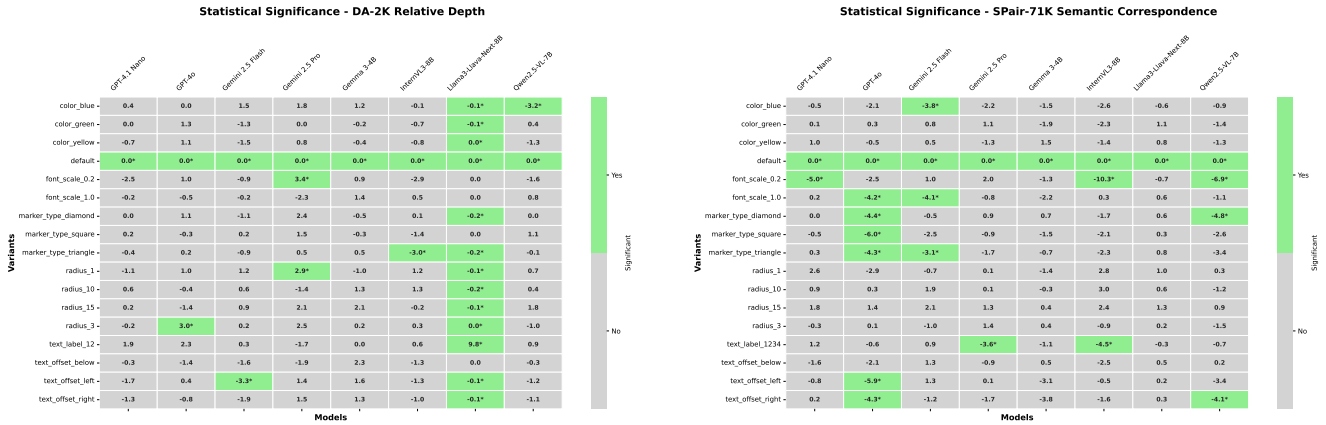
B. Change in accuracies on all marker styles

Below we show the full change in accuracy and rank for all marker styles across BLINK Relative Depth (Fig. 14), BLINK Semantic Correspondence (Fig. 15), SPair-71k (Fig. 16) and DA2k (Fig. 17), with the induced changes in model ranking isolated in Figures 13a-13d.

BLINK Relative Depth and Semantic Correspondence. For BLINK relative depth, changing only the marker style (color, size, shape, or label layout) yields drastic accuracy shifts, sometimes up to roughly 15% for individual models, as shown in the accuracy heatmap in Fig. 14. These shifts are large enough to reorder nearby models in the leaderboard, with several mid-ranked systems moving up or down multiple positions across marker variants (Fig. 13a). BLINK semantic correspondence shows a similar pattern: accuracy often changes by more than 10% under different marker styles (Fig. 15), and these shifts again reorder models with similar default performance (Fig. 13b), so marker design alone can change which model appears to perform best on both BLINK tasks.

SPair-71k Semantic Correspondence. For SPair-71k semantic correspondence, changing the marker style shifts model accuracies in systematic ways, with some variants consistently helping or hurting broad groups of models, as shown in Fig. 16. These shifts are also large enough to change the relative ordering of mid-ranked systems, with multiple models swapping positions across marker styles in Fig. 13d.

DA2k Relative Depth. On DA2k relative depth, marker style changes also lead to clear accuracy shifts for most models, though typically smaller than on BLINK potentially due to its larger data size, as shown in Fig. 17. Variants that alter marker size or label layout tend to have the strongest effect, while pure color or shape changes are milder but still noticeable. These differences are often sufficient to reorder models that are close in performance, especially away from the very top of the leaderboard, as illustrated in the rank changes in Fig. 13c and the significance analysis in Fig. 10.



(a) DA2K (b) SPair71k

Figure 10. **Significance plots of marker variants.** Green indicates statistical significance under paired bootstrap. We see that each model has at least 1 marker variant which produces a statistically significant difference in accuracy compared to the default marker.

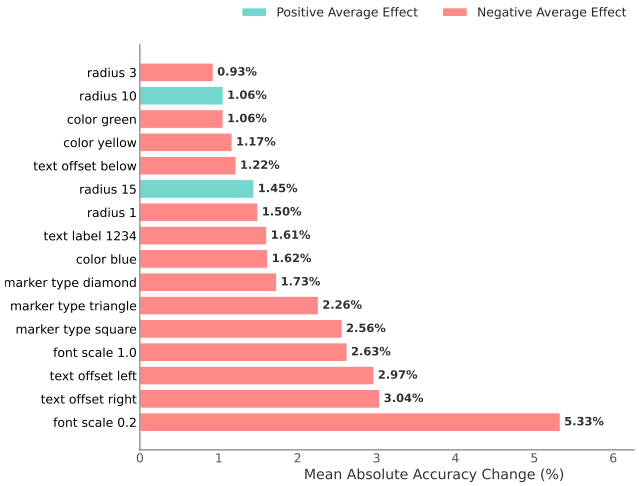
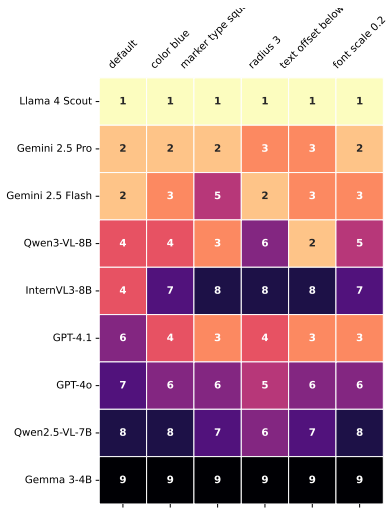
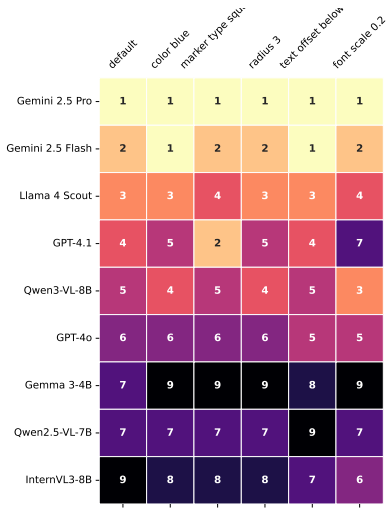


Figure 11. **Absolute marker impact.** Mean absolute accuracy shift on SPair induced by each marker variant. Variants which alter the text component of the visual marker typically result in the largest accuracy shifts.

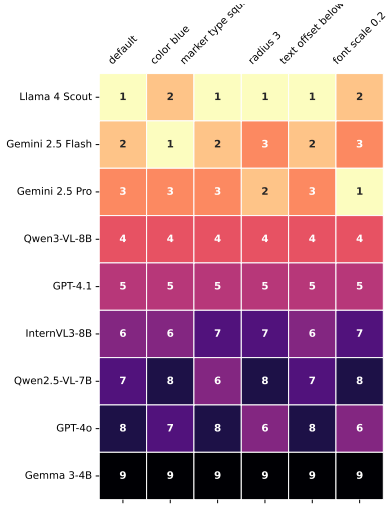
Figure 12. **Visual Marker Variants.** We explore 16 different visual markers in Section 5 of the main paper.



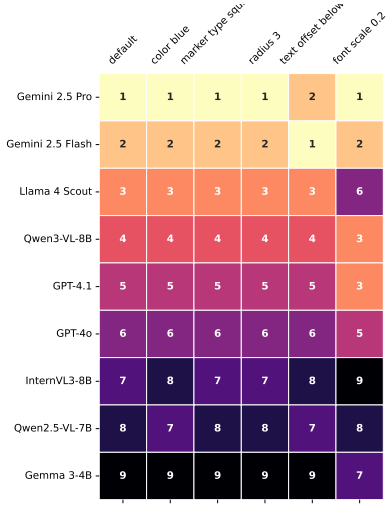
(a) BLINK RD



(b) BLINK SC



(c) DA2K



(d) SPair71k

Figure 13. **Change in rank for different marker styles** across BLINK RD, BLINK SC, DA2K, and SPair71k. For each dataset we see large fluctuations in rank across marker types, indicating that these tasks are highly sensitive to small visual changes.

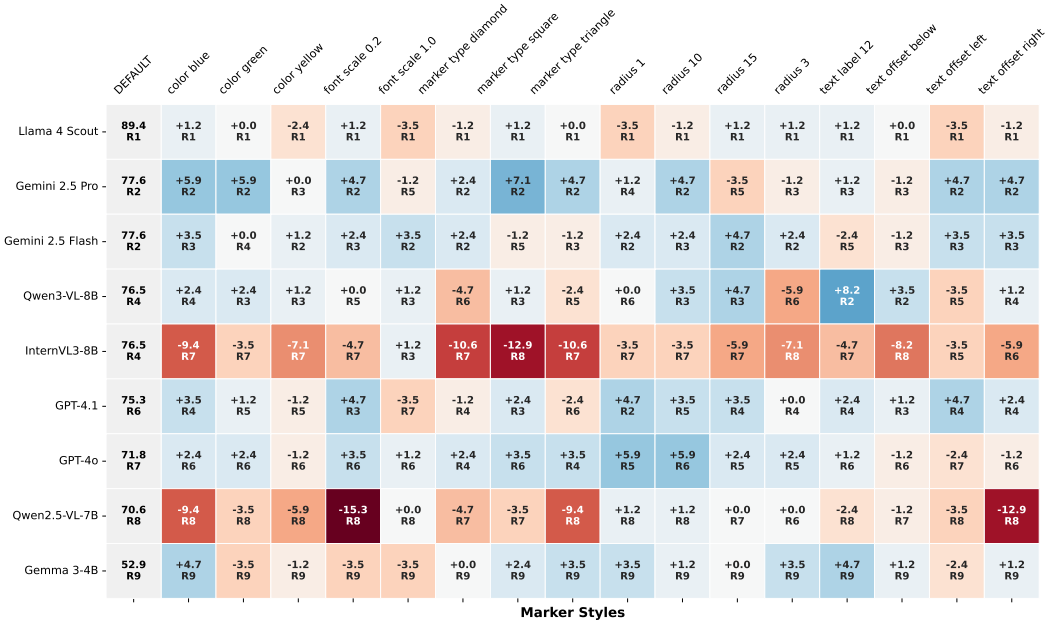


Figure 14. Change in accuracy and rank for different marker styles in BLINK relative depth task.

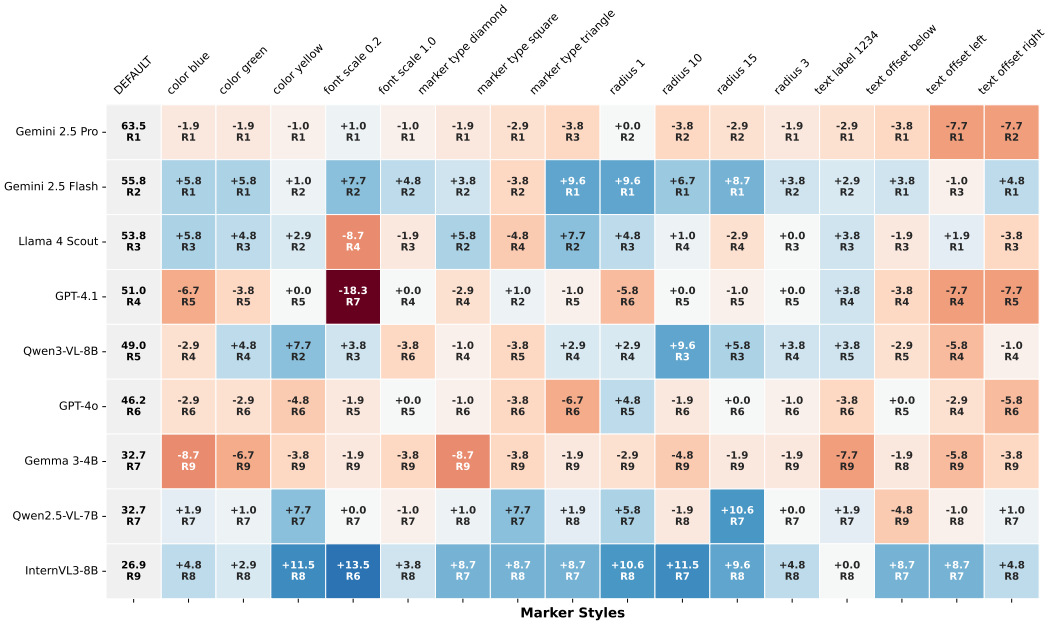


Figure 15. Change in accuracy and rank for different marker styles in BLINK semantic correspondence task.

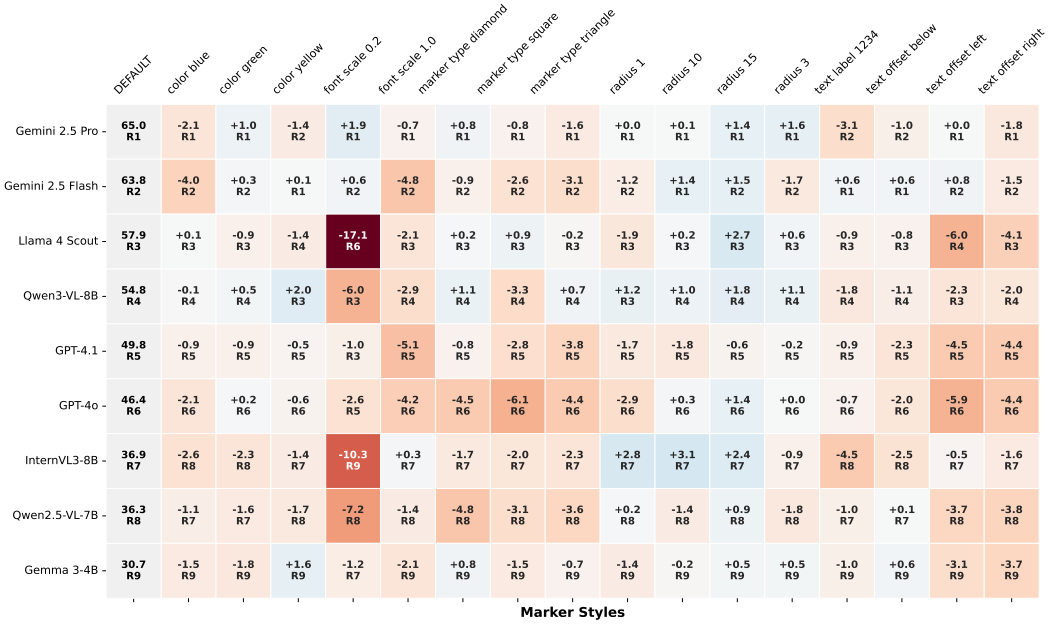


Figure 16. Change in accuracy and rank for different marker styles in SPair 71k.

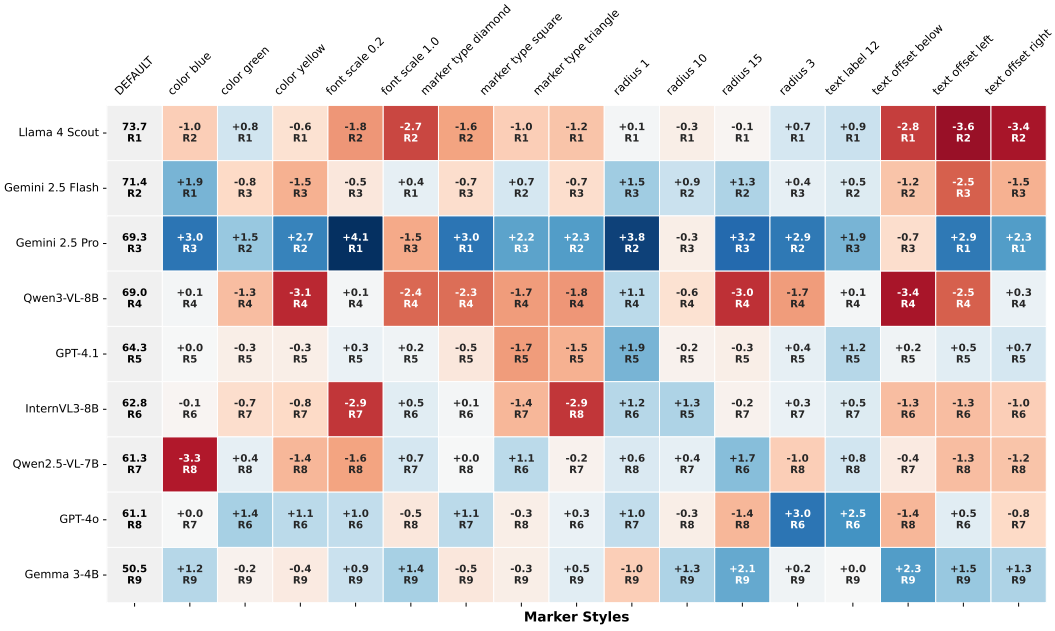


Figure 17. Change in accuracy and rank for different marker styles in DA2k.