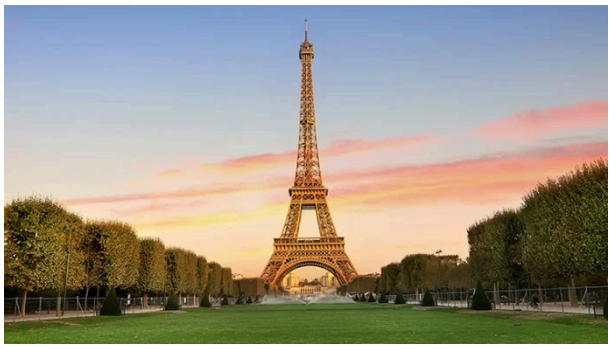
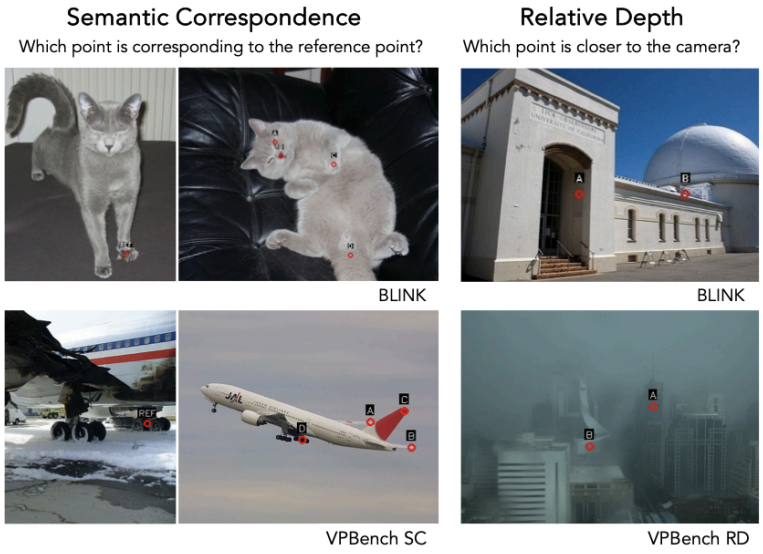
For those who are not up to date with the state of evaluation in computer vision, a key question that everyone wants to answer is how good is my model at seeing vs how good is my model at thinking. In other words, how can I measure a model's visual skills independent of its semantic knowledge? This is a harder task than you might think - many visual question answering samples rely heavily on world knowledge: if my model got the question below wrong, it's hard to determine if it is because it doesn't recognize this as a tower or because it doesn't know that this particular tower is in Paris.

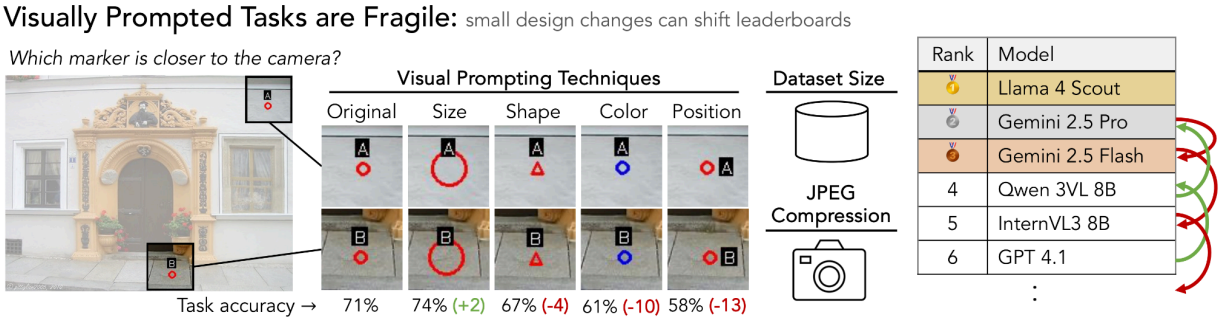


Question: in what city was this photo taken?

One such benchmark that aims to measure this pure visual capabilities is called BLINK, containing problems that humans can solve in "the blink of an eye", indicating that they require more visual processing than pure world knowledge. These tasks are things like determining which object in the image is closer to the camera or where objects are in relation to one another. Datasets like BLINK has incredibly high human performance but very low VLM performance, with some top models doing only slightly better than chance, and this has caused it to become a very popular benchmark for up in coming VLMs.



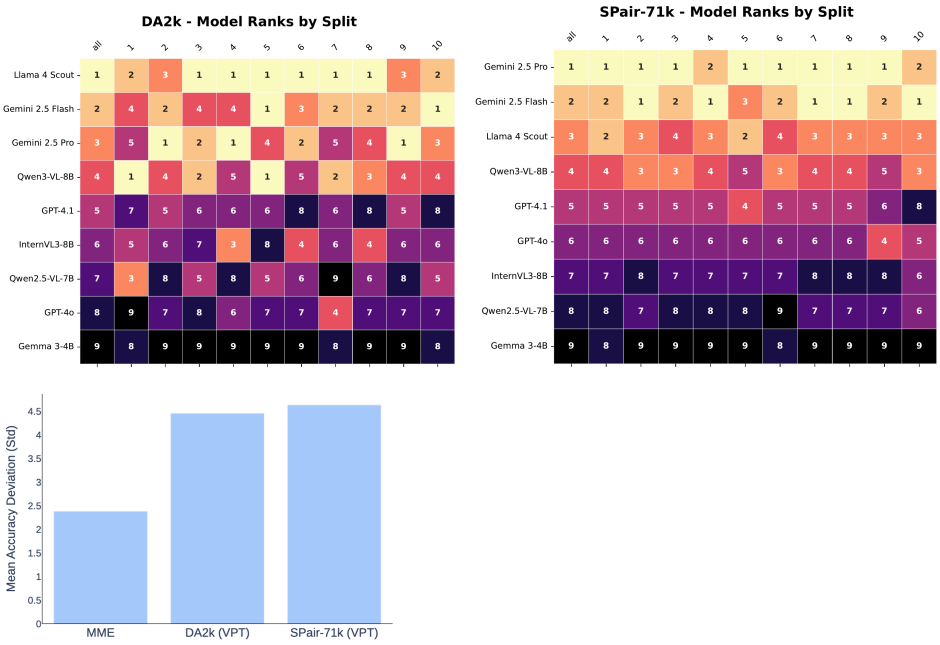
So what's the issue? Well several of these tasks are **visually prompted**, meaning a marker is placed within the image to assist in the task. While this altering of the image may seem inconsequential, we find that these visually prompted tasks are incredibly fragile: **simply changing the marker from red to blue can completely alter the leaderboard.**



In this post we will delve into the different sources of fragility for these types of benchmarks and provide some guidance on how we can improve upon them. While we use BLINK as an example of a popular dataset containing visually prompted tasks, this instability is seen across these sorts of visually prompted tasks *in general* (we <3 BLINK). To see how these changes affect leaderboards, we evaluate across 9 commonly used VLM's.

Ready? Let's jump in.

First let's get a glaring source of instability out of the way: dataset size. Each per-task dataset split of blink is 100-200 samples, which may seem like a reasonable amount, but since these tasks often get low accuracy (often only slightly above chance), the confidence intervals are huge. To show this, we created our own visually prompted benchmark VPBench, consisting of two tasks: relative depth and semantic correspondence. These datasets have around 10x the samples and thus much smaller confidence intervals. However, if we randomly create BLINK-sized data splits from VPBench, we *still* see that the model leaderboard across splits differs significantly. In fact, if you compare the average model accuracy change across these splits to the accuracy change of subsampling a non-visually prompted dataset like MME, we see that these visually prompted tasks are more sensitive to these smaller dataset sizes.



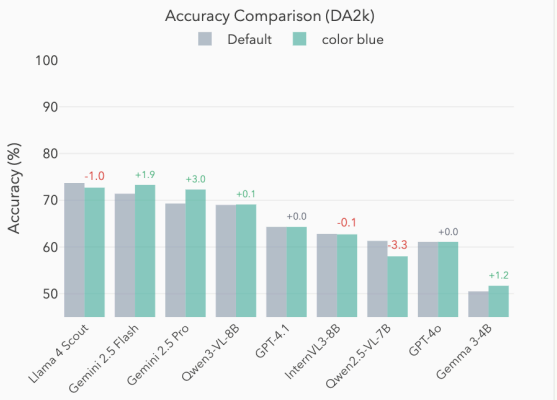
Next is marker style. This whole investigation actually started because we realized that within the BLINK dataset itself there are different visual markers: sometimes they are red, sometimes they are white, sometimes the text is above the circle, sometimes it's to the right. While this detail may seem completely inconsequential, we found huge accuracy shifts if you change the color, shape or size. In fact, we see that we can manipulate the leaderboard in our favor (or in someone else's disfavor) simply by strategically choosing a marker style.

Interactive marker comparison

Select a dataset and marker variant to see how accuracy and rankings change compared to the default.

Dataset: DA2k

Color: Blue | Marker Type: Square | Radius: 3 | Text Offset: Below | Font Scale: 0.2



MODEL	RANK	Δ RANK	SCORE
Llama 4 Scout	#2	↓ 1	72.7%
Gemini 2.5 Flash	#1	↑ 1	73.3%
Gemini 2.5 Pro	#3	– 0	72.3%
Qwen3-VL-8B	#4	– 0	69.1%
GPT-4.1	#5	– 0	64.3%
InternVL3-8B	#6	– 0	62.7%
Qwen2.5-VL-7B	#8	↓ 1	58.0%
GPT-4o	#7	↑ 1	61.1%
Gemma 3-4B	#9	– 0	51.7%

Small marker changes cause large, model-specific accuracy shifts and rank shuffles. Select a dataset and marker above to compare with the default evaluation.

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%

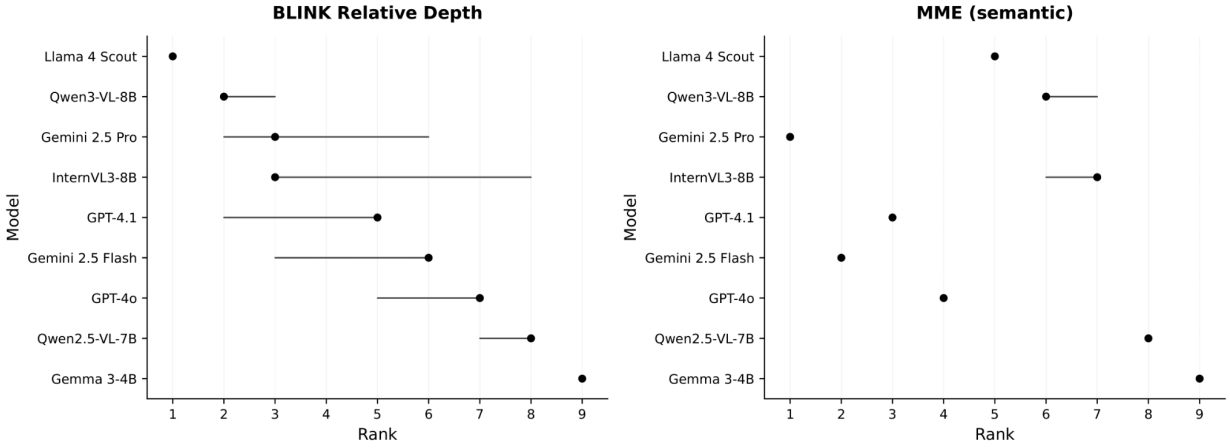
Inflate InternVL3-8B

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%

The last source of instability we also found by accident. When running our experiments we noticed that we were getting different results from each other, even when setting temperature to 0. Turns out the machines we were running on had different CUDA kernels or JPEG compression rates. These seemingly insignificant numerical differences are imperceptible to humans, yet they lead to noticeable changes in benchmark rankings.

Again we see that compared to a non visually prompted benchmark, visually prompted tasks see a large amount of variability from an imperceptible visual change.



With all these sources of instability, what can we do about it? Well larger dataset sizes help a lot: for instance, the ranking changes you see with JPEG compression decrease a lot when you increase your dataset size. A good rule of thumb is if your confidence intervals are wide: you need more eval data. To help with this we released our expanded visually prompted dataset VPBench, with ~10x the data.

More broadly, these results highlight a fundamental issue with visually prompted benchmarks: they often conflate visual reasoning ability with sensitivity to superficial cues. Without addressing this fragility, leaderboard movements may say more about marker design than about progress in visual understanding.