VB: Visibility Benchmark for Visibility and Perspective Reasoning in Images

Neil Tripathi*

February 24, 2026

Abstract

We present VB, a benchmark that tests whether vision-language models can determine what is and is not visible in a photograph—and abstain when a human viewer cannot reliably answer. Each item pairs a single photo with a short yes/no visibility claim; the model must output `VISIBLY_TRUE`, `VISIBLY_FALSE`, or `ABSTAIN`, together with a confidence score. Items are organized into 100 families using a 2×2 design that crosses a minimal image edit with a minimal text edit, yielding 300 headline evaluation cells. We score models on confidence-aware accuracy with abstention (CAA), minimal-edit flip rate (MEFR), confidence-ranked selective prediction (SelRank), and second-order perspective reasoning (ToMAcc). We evaluate nine models spanning flagship and prior-generation closed-source systems, and open-source models from 8B to 12B parameters. GPT-4o and Gemini 3.1 Pro tie for the best composite score (0.728 and 0.727), followed by Gemini 2.5 Pro (0.678). The best open-source model, Gemma 3 12B (0.505), surpasses one prior-generation closed-source system. Text-flip robustness exceeds image-flip robustness for six of nine models, and confidence calibration varies substantially—GPT-4o and Gemini 2.5 Pro achieve similar accuracy yet differ sharply in selective prediction quality.

1 Introduction

Vision-language models are increasingly deployed in settings where incorrect visual judgments carry real consequences—autonomous driving systems that must detect obscured pedestrians, assistive technologies that describe scenes to blind users, and medical imaging tools that flag ambiguous findings. In all of these domains, a model that guesses when the visual evidence is insufficient can be more dangerous than one that explicitly withholds judgment.

Visibility is a prerequisite for safe and reliable image-grounded reasoning. Many visually phrased questions are unanswerable from a single photo because the relevant evidence is occluded, out of frame, too small, too dark, or not visually observable at all. This paper introduces VB, a benchmark designed to quantify whether a system can:

- verify simple visibility claims from one image and a short question (testing whether models distinguish "visible" from "present"),
- respond appropriately to minimal edits that should flip the correct label (testing robustness to controlled perturbations),
- abstain when a human viewer cannot reliably answer from the photo (testing calibrated withholding of judgment).

*neiltripathi93@gmail.com

We additionally include a MULTI_AGENT / SECOND_ORDER slice that tests second-order perspective judgments—for example, whether the photo supports a claim about what one person can infer about another person's visual access. This capability is essential for collaborative and social reasoning grounded in images.

Contributions. This paper makes four contributions:

1. The VB benchmark: a task definition, an eight-category visibility taxonomy, and a 2×2 family design that crosses minimal image edits with minimal text edits over 100 families.
2. A metric suite—CAA, MEFR, SelRank, and a weighted composite—tailored to visibility reasoning with explicit abstention and confidence calibration.
3. An evaluation of nine vision-language models that reveals capability gaps between flagship and open-source systems, an asymmetry between text-flip and image-flip robustness, and substantial variation in confidence calibration.
4. Public release of the full dataset, per-image metadata, and evaluation infrastructure.

Data release. The full dataset, metadata, and evaluation code are released at:

https://github.com/neilt93/Paper-with-Davis

2 Qualitative examples

Figure 1 illustrates the 2×2 family construction. It reuses two images (I^0 and I^1) and two questions (q^0 and q^1) to yield four evaluated cells. If the referenced image files are not present locally, the document will display explicit placeholders instead of failing to compile.

Figure 2 shows one representative base/flip pair per primary category. Each pair uses the base question q^0 and displays the base image I^0 and the minimally edited image I^1 (the IMAGE_FLIP cell). For space, we omit the corresponding text-edited question q^1 and the DOUBLE_FLIP cell, but every family in the released benchmark includes all four cells (I^a, q^b) for $a, b \in \{0, 1\}$ (Section 4.5).

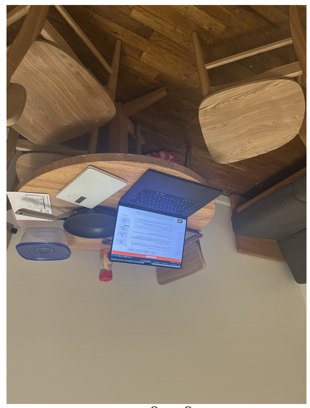
Note on `ABSTAIN` examples. `ABSTAIN` is a valid gold label in VB when a careful human cannot decide. The main headline results in Section 8 focus on the strict XOR headline subset used for score aggregation. `ABSTAIN`-labelled items are included for completeness and analysis, and we recommend reporting their separate accuracy and abstention rates in future releases.

3 Related work

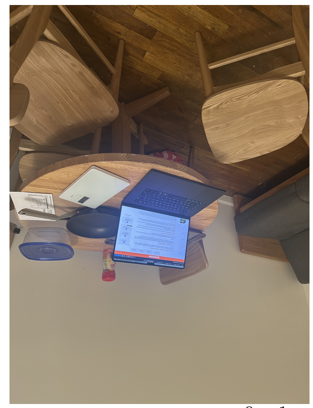
Unanswerable visual questions and withholding judgment. Davis points out that many questions about images are inherently unanswerable even with perfect vision because the relevant facts can be occluded, outside the frame, or non-visual [2]. Several benchmarks highlight that real images often do not support a definitive answer. VizWiz collects question–image pairs from blind or vision-impaired users, many of which lack sufficient visual evidence for a confident answer [7]. Recent work explicitly evaluates abstention on unanswerable visual questions, for example UNK-VQA [6] and TUBench [8]. VB differs from these efforts in two ways: it tests not just *whether* a question is unanswerable but *why* (via reason codes tied to specific visibility factors), and it uses controlled minimal edits to verify that models change their judgments when and only when the underlying evidence changes.

Figure 1: A single family shown in the full 2×2 layout. This figure uses two images and two questions, repeated across cells. To reproduce the figure, copy the corresponding files into `figs/examples/` using the filenames in the `\maybeincludegraphics` calls.

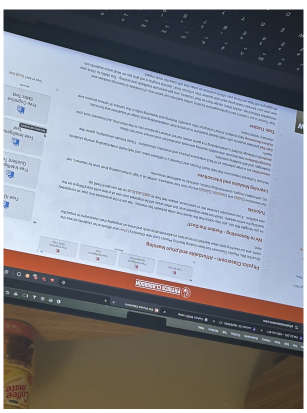
Family AV-04 (distance/readability).



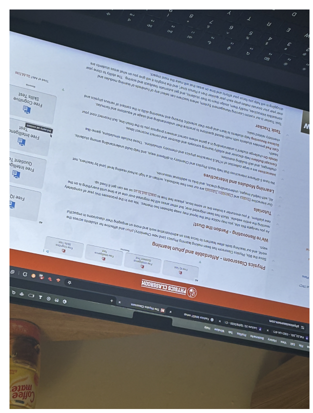
(a) **BASE** (I^0, q^0).
Q: "Can you read the small text on the laptop?"
Gold: `VISIBLY_FALSE`.



(b) **TEXT_FLIP** (I^0, q^1).
Q: "Can you not read the small text on the laptop?"
Gold: `VISIBLY_TRUE`.



(c) **IMAGE_FLIP** (I^1, q^0).
Gold: `VISIBLY_TRUE`.



(d) **DOUBLE_FLIP** (I^1, q^1).
Gold: `VISIBLY_FALSE`.

Figure 2: Representative base/flip examples (one family per primary category). Each pair shows the BASE cell (I^0, q^0) and the IMAGE_FLIP cell (I^1, q^0). Most shown pairs are from the strict XOR headline subset, where BASE is `VISIBLY_FALSE` and IMAGE_FLIP is `VISIBLY_TRUE` by construction (Section 4.5). The INSUFFICIENT_CONTEXT pair illustrates a case where the correct label is `ABSTAIN`.

(a) **LD-02 BASE.**
Q: "Is the sign text clearly readable in this photo?"
Gold: `VISIBLY_FALSE`.

(b) **LD-02 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.

(c) **OF-13 BASE.**
Q: "Is the microwave visible in this photo?"
Gold: `VISIBLY_FALSE`.

(d) **OF-13 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.

(e) **OC-11 BASE.**
Q: "Is the room number clearly visible in this photo?"
Gold: `VISIBLY_FALSE`.

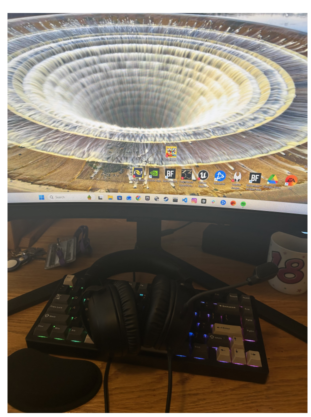
(f) **OC-11 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.

(g) **GD-06 BASE.**
Q: "Is he clearly looking at the laptop screen?"
Gold: `VISIBLY_FALSE`.

(h) **GD-06 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.

(i) **AV-01 BASE.**
Q: "Is the building number clearly readable in this photo?"
Gold: `VISIBLY_FALSE`.

(j) **AV-01 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.



(k) **NV-08 BASE.**
Q: "Is the song title/track info visible in this photo (screen/now playing)?"
Gold: `VISIBLY_FALSE`.

(l) **NV-08 IMAGE_FLIP.**
Gold: `VISIBLY_TRUE`.

Visibility factors: gaze, occlusion, and field of view. Datasets such as GazeFollow [12] and Following Gaze Across Views [13] study gaze targets and off-frame gaze. Work on occlusion reasoning analyzes how blockers and partial views affect recognition and localization [9]. VB differs by evaluating label-level support for a visibility claim under controlled edits rather than requiring precise localization.

Hallucination and faithfulness in vision-language models. Benchmarks such as MME [4] and POPE [11] evaluate multimodal perception, cognition, and hallucinated object claims. VB complements these by isolating a narrower but safety-relevant skill: deciding whether a claim is supported by visible evidence, and withholding judgment when it is not.

Selective prediction, reject options, and risk-coverage. Selective classification formalizes the trade-off between risk and coverage when a classifier may reject [3]. SelectiveNet and follow-up work evaluate selective prediction using risk-coverage curves and related summary statistics [5, 16]. VB uses a confidence-ranked selective prediction score to test whether model confidence aligns with correctness among answered items.

Second-order perspective and theory-of-mind style probes. Recent work motivates stress tests for second-order perspective judgments and their brittleness under structured perturbations [10, 15]. VB includes a dedicated MULTI_AGENT / SECOND_ORDER slice in which the label depends on what the photo supports about one agent's knowledge of another agent's visual access. Unlike text-only theory-of-mind probes, VB requires that second-order judgments be grounded in image evidence—the model must reason about what is visually accessible to each agent depicted in the photograph.

4 Benchmark design

4.1 Task definition and labels

Each item consists of an image and a short yes/no question that expresses a *visibility claim*. The system's job is to decide whether that claim is supported by visible evidence in the photo.

The label space is:

- `VISIBLY_TRUE`: the claim is supported by visible evidence, defined as "a careful human would answer *yes* from this photo with reasonable confidence";
- `VISIBLY_FALSE`: the claim is contradicted by the photo, defined as "a careful human would answer *no* from this photo with reasonable confidence";
- `ABSTAIN`: the photo does not support either a confident yes or a confident no.

The labels always refer to whether the *question's claim* is supported. This allows both positive and negative question forms, for example "Is the serial number readable?" versus "Is the serial number unreadable?"

4.2 Structured outputs

Systems output a single structured prediction:

- `label` $\in$ {`VISIBLY_TRUE`, `VISIBLY_FALSE`, `ABSTAIN`},

- `reason_code` $\in$ {GAZE_DIRECTION, OCCLUSION, OUT_OF_FRAME, LIGHTING_DISTANCE, INHERENTLY_NONVISUAL, AUGMENTED_VISION_REQUIRED, INSUFFICIENT_CONTEXT, MULTI_AGENT_SECOND_ORDER, NONE},
- `confidence` $\in [0, 1]$, interpreted as the model's probability that its chosen label is correct.

Reason-code conventions. If `label`=`VISIBLY_TRUE`, set `reason_code`=`NONE`. If `label`=`ABSTAIN`, choose exactly one limiting-factor code. If `label`=`VISIBLY_FALSE`, choose a limiting-factor code when the claim is false due to a visibility limitation (for example, "Is the serial number readable?" when the serial number is present but blurred). If the claim is directly refuted with no visibility limitation required (for example, "Is the mug not visible?" when the mug is plainly visible), set `reason_code`=`NONE`.

Why include `reason_code`? Reason codes make failures more actionable and more interpretable. They help distinguish "the evidence is missing" (out of frame) from "the evidence is present but blocked" (occlusion) or "present but too small" (distance), which correspond to different corrective actions (change viewpoint, remove blocker, move closer, increase illumination, or gather context).

Choosing a single code. If multiple limiting factors apply, we use the following precedence to select one:

```
OCCLUSION > OUT_OF_FRAME > GAZE_DIRECTION > LIGHTING_DISTANCE >
AUGMENTED_VISION_REQUIRED > INHERENTLY_NONVISUAL > INSUFFICIENT_CONTEXT >
MULTI_AGENT_SECOND_ORDER
```

The same precedence is included in the evaluation prompt in Section 6.

4.3 Family structure: 2×2 design

Items are grouped into families. Each family is built around:

- a base image I^0 and an edited image I^1 that differs by one atomic scene change,
- a base question q^0 and a text-edited question q^1 that flips the underlying claim being tested.

This yields four evaluated cells (I^a, q^b) for $a, b \in \{0, 1\}$:

$$(I^0, q^0),\ (I^0, q^1),\ (I^1, q^0),\ (I^1, q^1).$$

4.4 Atomic scene changes (minimal image edits)

An *atomic scene change* is a minimal change to the photographed world intended to affect exactly one visibility factor relevant to the family claim, while leaving the remainder of the scene as stable as possible. In practice, families were re-shot rather than digitally edited, and the atomic change was enacted physically (for example, moving closer to a screen, shifting an occluder, or moving a target slightly into frame).

We aim for the following constraints:

- **Single causal factor:** the edit targets one visibility factor (for example occlusion or distance) for the referent used in the question.
- **No incidental cues:** no new salient objects are introduced unless they are the edit itself (for example the occluder being moved).
- **Stable viewpoint:** camera position and framing are held fixed unless the family tests OUT_OF_FRAME or LIGHTING_DISTANCE and requires a controlled viewpoint change.

Table 1: Primary category label on each scene, what it tests, an example prompt (written to avoid encoding the category in the question text), and the number of base pictures in the current release.

Category (primary label on scene)	**What it tests**	**Example prompt (short)**	**# Base Pictures**
GAZE_DIRECTION	Head and eye orientation versus target	"Is Pat looking at the mug?"	20
OCCLUSION	Opaque blockers along line of sight	"Is the metal key blade visible?"	20
OUT_OF_FRAME	Target outside current view or crop	"Is the dog visible in the photo?"	15
LIGHTING_DISTANCE	Darkness, glare, or distance limits	"Is the sign text readable?"	13
INHERENTLY_NONVISUAL	Properties vision cannot reveal	"Is the device PIN visible anywhere?"	11
AUGMENTED_VISION_REQUIRED	Requires magnification not available from the base photo	"Is the fine print readable?"	7
INSUFFICIENT_CONTEXT	Under-specified referents or missing scene facts	"Is the correct key visible?"	7
MULTI_AGENT / SECOND_ORDER	Second-order perspective judgments grounded in one photo	"Does Bob know Alice cannot see the card?"	7
TOTAL			100

- **Stable lighting:** lighting conditions are held fixed unless the family targets LIGHTING_DISTANCE.

The DOUBLE_FLIP cell (Section 4.5) provides a diagnostic for unintended interactions, since composing the text and image edits should re-invert the claim under the intended construction.

4.5 XOR construction and the diagnostic fourth cell

For the strict headline subset used in the main score, families are constructed so that the gold labels follow a fixed XOR pattern over the two edits:

$$y^{00} = \texttt{VISIBLY\_FALSE}, \quad y^{01} = \texttt{VISIBLY\_TRUE}, \quad y^{10} = \texttt{VISIBLY\_TRUE}, \quad y^{11} = \texttt{VISIBLY\_FALSE}.$$

Intuitively, the base cell is designed to be confidently refuted by the photo, while either a text edit alone or an image edit alone makes the claim supported. When both edits are applied, the claim is again refuted.

We treat the first three cells as **headline cells** for the main score:

$$(I^0, q^0)\ \text{BASE}, \quad (I^0, q^1)\ \text{TEXT\_FLIP}, \quad (I^1, q^0)\ \text{IMAGE\_FLIP}.$$

The fourth cell (I^1, q^1) (**DOUBLE_FLIP**) is **diagnostic only**. We report it separately and do not include it in the composite headline score.

4.6 Primary categories

VB is organized around mutually exclusive primary visibility factors at the family level. Each base image is tagged with exactly one primary category.

4.7 Multi-agent and second-order visibility

The MULTI_AGENT / SECOND_ORDER subset targets second-order judgments, where the system must decide whether the photo supports a claim about one agent's knowledge of another agent's visual access.

Writing rule (referents must be clear). SECOND_ORDER questions name agents and targets explicitly (for example "Bob", "Alice", and "the card"), rather than relying on pronouns. If the referent would be unclear to a careful human, the intended gold label is `ABSTAIN` (reason code `INSUFFICIENT_CONTEXT`).

Construction principle. SECOND_ORDER families follow the same 2×2 structure. For the strict headline subset, we use the XOR pattern in Section 4.5. For additional `ABSTAIN`-labelled families, we require that independent annotators agree that a careful human cannot answer from the photo.

4.8 Data collection

All images were collected by the authors in and around the NYU campus. Indoor scenes were photographed in student dormitory spaces (bedrooms, kitchens, laundry rooms, corridors, stairwells, and common areas). Outdoor scenes were photographed in the surrounding city (pavements, street crossings, car parks, and neighbourhood parks). Images were captured at approximately human eye height using a smartphone camera. The release includes per-image metadata and editing provenance.

5 Evaluation and scoring

5.1 Gold labels and cells

For the strict XOR headline subset, gold labels are fixed by construction (Section 4.5). We refer to the four cells as:

- `BASE`: (I^0, q^0) with gold `VISIBLY_FALSE`
- `TEXT_FLIP`: (I^0, q^1) with gold `VISIBLY_TRUE`
- `IMAGE_FLIP`: (I^1, q^0) with gold `VISIBLY_TRUE`
- `DOUBLE_FLIP`: (I^1, q^1) with gold `VISIBLY_FALSE` (diagnostic)

Models may output `ABSTAIN` when they cannot decide with reasonable confidence from the image and question.

5.2 Confidence-aware accuracy with abstention (CAA)

Let item i have gold label $y_i \in \{\texttt{VISIBLY\_TRUE}, \texttt{VISIBLY\_FALSE}\}$. A model outputs $\hat{y}_i \in \{\texttt{VISIBLY\_TRUE}, \texttt{VISIBLY\_FALSE}, \texttt{ABSTAIN}\}$ and a confidence $\hat{c}_i \in [0, 1]$. (For evaluation, $\hat{c}_i$ is used only when $\hat{y}_i \neq \texttt{ABSTAIN}$.)

We define confidence-aware accuracy with abstention (CAA) using a partial credit parameter $\alpha \in [0, 1]$:

$$\text{score}_i = \begin{cases} \alpha & \text{if } \hat{y}_i = \texttt{ABSTAIN}, \\ \hat{c}_i & \text{if } \hat{y}_i = y_i, \\ 0 & \text{if } \hat{y}_i \neq y_i \text{ and } \hat{y}_i \neq \texttt{ABSTAIN}. \end{cases}$$

Then

$$\mathrm{CAA} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{score}_i.$$

Justification. CAA rewards high confidence when correct, gives zero credit for incorrect answers regardless of confidence, and gives fixed partial credit for abstention. This avoids a gaming vulnerability where low-confidence wrong answers could score well under alternative formulations. The design is inspired by the selective prediction literature, where models trade coverage for reduced risk [3, 5]. We use $\alpha = 0.25$ by default to give abstention a small but non-trivial value, reflecting that withholding judgment can be preferable to a guess in safety-relevant settings.

Headline CAA is computed on the three headline cells only (`BASE`, `TEXT_FLIP`, `IMAGE_FLIP`).

5.3 Minimal edit flip rates (MEFR)

We measure sensitivity to minimal edits along each axis, conditioning on correctness of the base cell. For correctness checks, `ABSTAIN` counts as incorrect.

Let $c_f^{ab} = \mathbf{1}[\hat{y}_f^{ab} = y_f^{ab}]$ for family f and cell (a, b).

$$\mathrm{I\_MEFR} = \frac{\sum_{f=1}^{F} \mathbf{1}\left[c_f^{00} = 1 \wedge c_f^{10} = 1\right]}{\sum_{f=1}^{F} \mathbf{1}\left[c_f^{00} = 1\right]}, \quad \mathrm{T\_MEFR} = \frac{\sum_{f=1}^{F} \mathbf{1}\left[c_f^{00} = 1 \wedge c_f^{01} = 1\right]}{\sum_{f=1}^{F} \mathbf{1}\left[c_f^{00} = 1\right]}.$$

$$\mathrm{MEFR} = \tfrac{1}{2}(\mathrm{I\_MEFR} + \mathrm{T\_MEFR}).$$

Denominators. Because MEFR conditions on correctness of the BASE cell, the effective denominator (the number of families with correct BASE predictions) can vary across models, especially when abstention is frequent. We report MEFR denominators alongside selective prediction diagnostics in Table 3.

5.4 Confidence-ranked selective prediction score (SelRank)

We evaluate selective prediction using confidence-ranked answering on headline cells. We consider only *answered* predictions (`VISIBLY_TRUE` or `VISIBLY_FALSE`) and exclude `ABSTAIN` from coverage.

Sort answered items by confidence in descending order. For each prefix length k, compute coverage $\mathrm{cov}(k) = k/n$ and answered accuracy $\mathrm{acc}(k)$. Let A_{model} be the area under the answered accuracy versus coverage curve (trapezoidal rule). Let p be the overall answered accuracy (the flat baseline).

This is closely related to the standard area under the risk-coverage curve (AURC) used in selective classification [3, 16], since $\mathrm{risk}(k) = 1 - \mathrm{acc}(k)$. We use a normalized "gain" formulation where higher is better (unlike standard AURC where lower is better), and name it SelRank to avoid confusion.

We report a normalized score (upper-capped at 1):

$$\mathrm{SelRank} = \min\left(1,\ \frac{A_{\mathrm{model}} - p}{1 - p}\right).$$

This value is 0 when confidence ranking is no better than the flat baseline, and increases when higher confidence corresponds to higher correctness. Negative values indicate that confidence ranking is *anti*-informative (lower-confidence answers tend to be more correct).

5.5 Multi-agent / second-order accuracy (ToMAcc)

On the MULTI_AGENT / SECOND_ORDER strict subset, we report:

$$\text{ToMAcc} = \frac{\text{number of correct predictions on SECOND\_ORDER items}}{\text{total number of SECOND\_ORDER items}}.$$

For ToMAcc, `ABSTAIN` counts as incorrect.

5.6 Double flip diagnostic accuracy (DFAcc)

We report diagnostic accuracy on the DOUBLE_FLIP cell:

$$\text{DFAcc} = \frac{\text{number of correct predictions on DOUBLE\_FLIP}}{\text{total number of DOUBLE\_FLIP items evaluated}}.$$

DFAcc is reported separately and is not included in the final composite score.

5.7 Final composite score

The final benchmark score combines four headline components with fixed weights:

$$\textsc{FinalScore} = 0.70 \cdot \text{CAA} + 0.15 \cdot \text{MEFR} + 0.10 \cdot \text{SelRank} + 0.05 \cdot \text{ToMAcc}.$$

Weight choice. These coefficients reflect a prioritization: correctness with calibrated abstention is the primary objective, robustness to minimal edits is next, confidence ranking for selective answering is third, and second-order performance is included as a small but distinct capability axis. We therefore report all component metrics and encourage alternative weightings for different application priorities.

If no SECOND_ORDER items are present in an evaluation run, weights are renormalized over the remaining terms.

6 Prompt template

We used a single, explicit instruction block (provided as the user message) to standardise how models interpret the task, handle ambiguity, and format outputs. Two parts of the prompt deserve emphasis. First, the referring-expression rules ensure that the target entity can be identified from the image plus the question text alone, without any hidden annotation channel (such as arrows or markings). Second, the instruction "Do not require arrows, circles, or any on-image markings" clarifies that the benchmark does not assume any visual markup; questions must be answerable from the unedited photo.

The prompt is shown verbatim below, with `{question}` substituted per item.

```
You are given one photo and one question. Decide whether the question's claim is true using
↪  only pixels in this photo. Do not guess and do not use world knowledge.

Observer

The observer is the camera unless the question explicitly names a person in the image as the
↪  observer.

How to interpret questions
```

```
Treat questions like "Is X visible/readable/clear?" or "Is person A looking at X?" as a true/false claim you must judge from the image.

For open-ended content questions (eg "what's in the box?", "what time is it?"), convert them into: "Is that information visually determinable from this photo right now?"

Referring expressions (who is "he/she/they/it")

Do not require arrows, circles, or any on-image markings.

If the question uses pronouns or vague references, you must be able to identify the referent unambiguously from the image plus the question text alone.

Acceptable disambiguation: only one plausible referent exists, or the question uniquely specifies it via clear visual descriptors (eg "the man on the left", "the woman in the red jacket", "the cat on the sofa").

If multiple referents are plausible and the text does not uniquely pick one, choose ABSTAIN with reason_code=INSUFFICIENT_CONTEXT.

Output format
Return exactly one minified JSON object (no prose, no markdown, no trailing commas) with keys in this order:
{"label":"VISIBLY_TRUE|VISIBLY_FALSE|ABSTAIN","reason_code":"GAZE_DIRECTION|OCCLUSION|OUT_OF_FRAME|LIGHTING_DISTANCE|INHERENTLY_NONVISUAL|AUGMENTED_VISION_REQUIRED|INSUFFICIENT_CONTEXT|MULTI_AGENT_SECOND_ORDER|NONE","confidence":number}

Label meanings

VISIBLY_TRUE: the claim in the question is clearly true from pixels.

VISIBLY_FALSE: the claim in the question is clearly false from pixels.

ABSTAIN: you cannot decide true vs false with reasonable confidence from this image.

Reason codes

If label="VISIBLY_TRUE", set reason_code="NONE".

If label="ABSTAIN", pick exactly one reason_code explaining what prevents a decision.

If label="VISIBLY_FALSE":

If the claim is false because the opposite is clearly true (eg the question asserts "not visible" but it is plainly visible), you may set reason_code="NONE".

Otherwise pick exactly one limiting-factor reason_code explaining why the claim is false.

Precedence if multiple apply:
OCCLUSION > OUT_OF_FRAME > GAZE_DIRECTION > LIGHTING_DISTANCE > AUGMENTED_VISION_REQUIRED > INHERENTLY_NONVISUAL > INSUFFICIENT_CONTEXT > MULTI_AGENT_SECOND_ORDER

Transparent clear glass is non-occluding; frosted/translucent counts as occluding for recognition.

Confidence

Confidence is your probability that your chosen label is correct (0.0 to 1.0).
```

```
Use your own internal thresholding. If you cannot decide with reasonable confidence, choose
↪  ABSTAIN.

Question: {question}
```

7 Experimental setup

We evaluated nine vision-language models on VB, organized into three groups:

- **Flagship closed-source (3):** Gemini 3.1 Pro (Google), GPT-5 (OpenAI), Claude Opus 4.5 (Anthropic)
- **Prior-generation closed-source (3):** GPT-4o (OpenAI), Gemini 2.5 Pro (Google), Claude 3.7 Sonnet (Anthropic)
- **Open-source 8–12B (3):** Gemma 3 12B [14], InternVL3-8B [17], Qwen3-VL-8B [1]

All models were queried with the prompt in Section 6. For each family, we ran the four cells (BASE, TEXT_FLIP, IMAGE_FLIP, DOUBLE_FLIP). Headline metrics used the first three cells only, yielding $3F = 300$ scored headline items per model. We used $\alpha = 0.25$ for CAA. SelRank was computed on answered items only, with `ABSTAIN` excluded from coverage.

Open-source models (Gemma 3 12B, InternVL3-8B, Qwen3-VL-8B) were run on a single NVIDIA RTX 3090 GPU (24 GB VRAM) using RunPod cloud infrastructure.

8 Results

8.1 Overall performance

Table 2 reports the main results on the strict headline subset for all nine models. GPT-4o and Gemini 3.1 Pro effectively tie for the highest overall FinalScore (0.728 and 0.727 respectively), followed by Gemini 2.5 Pro (0.678) and GPT-5 (0.625). Claude Opus 4.5 (0.570) forms a middle tier among flagship systems. Among prior-generation models, GPT-4o (0.728) substantially outperforms its tier peers Gemini 2.5 Pro (0.678) and Claude 3.7 Sonnet (0.476). The best open-source model, Gemma 3 12B (0.505), surpasses Claude 3.7 Sonnet, demonstrating that current open-source models at the 8–12B scale can exceed prior-generation closed-source systems on visibility reasoning. InternVL3-8B (0.445) and Qwen3-VL-8B (0.419) occupy the lower range.

Abstention behavior varies substantially across models. GPT-5 abstains most frequently (78 of 300 headline items), followed by Qwen3-VL-8B (50), while Gemini 3.1 Pro abstains least (14 of 300). Qwen3-VL-8B also produced 66 unparsable outputs across all cells (49 in headline cells), reducing its effective coverage. Because CAA awards partial credit for abstention, we report abstention counts alongside all headline metrics.

8.2 Minimal edit sensitivity

T_MEFR exceeds I_MEFR for six of nine models, indicating that text-flipped questions are generally handled more reliably than minimal image edits. This asymmetry is most pronounced for GPT-4o (T_MEFR 0.893 vs. I_MEFR 0.800) and Gemini 2.5 Pro (0.892 vs. 0.743). Three models reverse the pattern: InternVL3-8B (I_MEFR 0.610 vs. T_MEFR 0.373), Qwen3-VL-8B (0.307 vs. 0.180), and Claude 3.7 Sonnet (0.507 vs. 0.493), though the last reversal is marginal.

GPT-4o achieves the highest overall MEFR (0.847), followed by GPT-5 (0.819) and Gemini 2.5 Pro (0.818), indicating robust handling of both edit types. Qwen3-VL-8B shows the lowest MEFR (0.243), partly due to its low T_MEFR (0.180), suggesting difficulty with negated questions.

Table 2: Headline results on VB (300 scored headline items per model). CAA uses $\alpha = 0.25$. SelRank is a normalized selective ranking score on answered items (negative values indicate anti-informative confidence ranking). ABS is the number of abstentions among the 300 headline items. Models are grouped into three tiers (3/3/3): flagship closed-source, prior-generation closed-source, and open-source 8–12B.

Model	Final	CAA	I_MEFR	T_MEFR	MEFR	SelRank	ToMAcc	ABS
Gemini 3.1 Pro	0.727	0.760	0.659	0.871	0.765	0.394	0.810	14
GPT-5	0.625	0.622	0.793	0.845	0.819	0.237	0.857	78
Claude Opus 4.5	0.570	0.580	0.570	0.671	0.620	0.374	0.667	18
GPT-4o	0.728	0.769	0.800	0.893	0.847	0.144	0.952	16
Gemini 2.5 Pro	0.678	0.747	0.743	0.892	0.818	−0.106	0.857	18
Claude 3.7 Sonnet	0.476	0.508	0.507	0.493	0.500	0.192	0.524	30
Gemma 3 12B	0.505	0.543	0.424	0.644	0.534	0.087	0.714	25
InternVL3-8B	0.445	0.498	0.610	0.373	0.492	0.018	0.429	24
Qwen3-VL-8B	0.419	0.509	0.307	0.180	0.243	0.033	0.450	50

Table 3: Selective prediction and MEFR diagnostics on headline items. **Answered** counts non-ABSTAIN outputs among 300 headline items (for Qwen3-VL-8B, 49 headline items were unparsable and excluded). **Coverage** is answered/300. **AnsAcc** is accuracy on answered items only. **SelRank$_{\text{raw}}$** is the normalized selective ranking score before capping. **MEFR d** is the number of families with correct BASE prediction. Model grouping follows Table 2.

Model	Answered	Cov (%)	AnsAcc	SelRank$_{\text{raw}}$	MEFR d
Gemini 3.1 Pro	286	95.3	0.804	0.394	85
GPT-5	222	74.0	0.851	0.237	58
Claude Opus 4.5	282	94.0	0.684	0.374	79
GPT-4o	284	94.7	0.831	0.144	75
Gemini 2.5 Pro	282	94.0	0.794	−0.106	74
Claude 3.7 Sonnet	270	90.0	0.589	0.192	73
Gemma 3 12B	275	91.7	0.618	0.087	59
InternVL3-8B	273	91.0	0.553	0.018	59
Qwen3-VL-8B	201	67.0	0.582	0.033	72

8.3 Abstention, coverage, and confidence ranking

Since SelRank is answered-only, it is useful to also report answered coverage and answered accuracy. Table 3 reports answered fraction, answered-only accuracy, the MEFR denominator (families where BASE is correct), and the raw normalized SelRank.

Several patterns emerge. First, GPT-5 achieves the highest answered-only accuracy (0.851) but with unusually low coverage (74.0%, 78 abstentions), indicating a cautious strategy that substantially reduces its composite score. Second, GPT-4o achieves the second-highest answered accuracy (0.831) with high coverage (94.7%) and moderate calibration (SelRank 0.144). Third, Gemini 3.1 Pro achieves the best confidence calibration (SelRank 0.394) alongside strong answered accuracy (0.804, 95.3% coverage), meaning higher-confidence answers tend to be more accurate. By contrast, Gemini 2.5 Pro achieves similar answered accuracy (0.794) and coverage (94.0%) but has anti-informative confidence ranking (SelRank −0.106), meaning its lower-confidence answers tend to be more correct. Among open-source models, Qwen3-VL-8B shows the lowest effective coverage (67.0%) due to 49 unparsable headline outputs, though its answered-only accuracy (0.582) exceeds that of InternVL3-8B

(0.553).

8.4 Multi-agent and second-order subset

ToMAcc is computed on the MULTI_AGENT / SECOND_ORDER slice (7 families, 21 headline items). GPT-4o achieves the best performance on this subset (0.952), followed by GPT-5 and Gemini 2.5 Pro (both 0.857) and Gemini 3.1 Pro (0.810). Among open-source models, Gemma 3 12B scores 0.714, while Claude 3.7 Sonnet (0.524), Qwen3-VL-8B (0.450), and InternVL3-8B (0.429) perform near chance level. Claude Opus 4.5 scores 0.667. The strong performance of GPT-4o on second-order reasoning is notable, with all three prior-generation and flagship Google and OpenAI models substantially outperforming open-source alternatives.

Confidence intervals. With only 21 items, ToMAcc estimates have wide confidence intervals. Using Wilson's method, the 95% CI for ToMAcc = 0.952 (20/21) is approximately [0.77, 0.99], and for ToMAcc = 0.429 (9/21) is approximately [0.24, 0.64]. Rankings within this subset should be interpreted cautiously; expanding the SECOND_ORDER slice is a priority for future dataset releases.

8.5 Diagnostic: double flip

We also record DOUBLE_FLIP behavior (I^1, q^1) for diagnosis, but it is not included in FinalScore. The DOUBLE_FLIP cell can identify families where composing edits yields unexpected interactions, or where an intended atomic edit incidentally changes additional visibility factors. In this release we treat DOUBLE_FLIP primarily as a diagnostic signal and leave systematic family auditing based on DOUBLE_FLIP to future work.

8.6 Open-source versus closed-source

A substantial performance gap exists between flagship closed-source models and open-source models at the 8–12B scale. GPT-4o (0.728) and Gemini 3.1 Pro (0.727) outperform the best open-source model, Gemma 3 12B (0.505), by over 0.22 in composite score (roughly 30% relative). The gap is particularly pronounced on second-order reasoning (ToMAcc), where GPT-4o (0.952) and Gemini 3.1 Pro (0.810) far exceed all open-source alternatives (Gemma 3 12B: 0.714, InternVL3-8B: 0.429, Qwen3-VL-8B: 0.450).

However, the open-source tier shows meaningful differentiation. Gemma 3 12B (0.505) surpasses one prior-generation closed-source system—Claude 3.7 Sonnet (0.476)—demonstrating that visibility reasoning is not exclusively a closed-source capability. At the other extreme, Qwen3-VL-8B shows the lowest MEFR (0.243) and produced 66 unparsable outputs across all cells, suggesting output-format compliance issues that limit effective evaluation. InternVL3-8B (0.445) falls between its open-source peers, with notably high image-flip robustness (I_MEFR 0.610) but weak text-flip handling (T_MEFR 0.373).

9 Discussion

The evaluation of nine models on VB reveals three findings about the state of visibility reasoning in current vision-language models.

Flagship versus open-source gap. The top closed-source models (GPT-4o, Gemini 3.1 Pro) outperform the best open-source model (Gemma 3 12B) by roughly 30% in relative composite score (0.728 vs. 0.505). The gap is largest on second-order reasoning, where GPT-4o (ToMAcc 0.952) and Gemini 3.1 Pro (0.810) substantially exceed open-source alternatives. However, Gemma 3 12B surpasses one prior-generation closed-source system, suggesting that the gap is narrowing at the 8–12B parameter scale.

Text-flip versus image-flip asymmetry. T_MEFR exceeds I_MEFR for six of nine models, suggesting that models are generally better at tracking logical negation in text than at detecting subtle visual changes between minimally edited photos. This asymmetry has practical implications: text augmentation (rephrasing questions, testing negations) may be a more effective robustness intervention than visual augmentation for current systems. Three models reverse the pattern—InternVL3-8B, Qwen3-VL-8B, and Claude 3.7 Sonnet—though the last reversal is marginal (0.507 vs. 0.493). Qwen3-VL-8B shows the lowest T_MEFR (0.180), suggesting a floor effect in which some models cannot reliably handle negated questions.

Calibration variability. GPT-4o and Gemini 2.5 Pro achieve similar accuracy (CAA 0.769 vs. 0.747) yet differ sharply in selective prediction quality (SelRank 0.144 vs. −0.106). Gemini 3.1 Pro achieves the best calibration overall (SelRank 0.394) alongside strong accuracy (CAA 0.760). This dissociation implies that accuracy alone is insufficient for deployment decisions: a system whose confidence scores do not rank correctness cannot safely defer uncertain predictions. The connection to selective prediction literature [3, 5] is direct—SelRank measures exactly the property needed for risk-coverage trade-offs in practice.

Limitations. The current release has several limitations. First, 100 families provide limited statistical power for fine-grained comparisons, particularly on the 21-item SECOND_ORDER slice. Second, all images were collected in a narrow set of environments (campus buildings and nearby streets), so results may not generalize to other settings such as rural, industrial, or indoor-medical scenes. Third, the composite score uses fixed weights; we have not yet conducted a sensitivity analysis over alternative weightings. Fourth, the benchmark uses short yes/no questions to tightly control claims, which improves interpretability but does not cover longer multi-step reasoning. Fifth, Qwen3-VL-8B produced 66 unparsable outputs, reducing its effective evaluation set and potentially underestimating its capabilities.

Implications. Visibility reasoning is a prerequisite for safe image-grounded behavior, and the patterns observed here suggest three practical takeaways. First, the calibration gap between GPT-4o and Gemini 2.5 Pro indicates that systems should be evaluated not only on correctness but also on whether they express uncertainty in ways that support selective answering. Second, the T_MEFR > I_MEFR asymmetry (present in six of nine models) suggests that text augmentation (e.g., testing negated question variants) may be a more effective robustness intervention than visual augmentation for current models. Third, the low T_MEFR observed for Qwen3-VL-8B (0.180) indicates that some open-source models at the 8B scale still lack reliable negation handling, a prerequisite for the abstention-aware reasoning VB targets.

10 Conclusion

We introduced VB, a benchmark for visibility and perspective reasoning in images with a controlled 2×2 family construction, explicit abstention, and a dedicated MULTI_AGENT / SECOND_ORDER slice. We evaluate nine vision-language models across 100 families. GPT-4o and Gemini 3.1 Pro achieve the highest composite scores, while the best open-source model (Gemma 3 12B) surpasses one prior-generation closed-source system, demonstrating that visibility reasoning capabilities are beginning to transfer to the open-source tier at the 8–12B scale. Future work will expand the benchmark with additional families, a larger SECOND_ORDER slice, and images from broader environments to strengthen statistical power and generalizability.

Acknowledgements

I thank Ernest Davis for detailed feedback on benchmark design, paper presentation, and evaluation methodology.

References

[1] Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint*, arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

[2] Ernest Davis. Unanswerable questions about images and texts. *Frontiers in Artificial Intelligence*, 3:51, 2020. doi: 10.3389/frai.2020.00051. URL https://doi.org/10.3389/frai.2020.00051.

[3] Ran El-Yaniv. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(May):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint*, arXiv:2306.13394, 2023. URL https://arxiv.org/abs/2306.13394.

[5] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2151–2159. PMLR, 2019.

[6] Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan S. Kankanhalli. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *arXiv preprint*, arXiv:2310.10942, 2023. URL https://arxiv.org/abs/2310.10942.

[7] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, 2018.

[8] Xingwei He, Qianru Zhang, A-Long Jin, Yuan Yuan, and Siu-Ming Yiu. Tubench: Benchmarking large vision-language models on trustworthiness with unanswerable questions. *arXiv preprint*, arXiv:2410.04107, 2024. URL https://arxiv.org/abs/2410.04107.

[9] Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, and Manmohan Chandraker. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–121, 2020.

[10] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint*, arXiv:2302.02083, 2023. URL https://arxiv.org/abs/2302.02083.

[11] Yifan Li et al. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*, 2023. URL https://arxiv.org/abs/2305.10355. Also available as arXiv:2305.10355.

[12] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 199–207, 2015.

[13] Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze across views. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] Gemma Team. Gemma 3 technical report. *arXiv preprint*, arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

[15] Tomer D. Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint*, arXiv:2302.08399, 2023. URL https://arxiv.org/abs/2302.08399.

[16] Neeraj Varshney, Swaroop Mishra, Pallavi Vijayakumar, and Chitta Baral. Investigating selective prediction approaches across nlp tasks. *Findings of the Association for Computational Linguistics*, 2022. URL https://aclanthology.org/2022.findings-acl.

[17] Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint*, arXiv:2504.10479, 2025. URL https://arxiv.org/abs/2504.10479.