VB: Visibility Benchmark for Visibility and Perspective Reasoning in Images

Neil Tripathi*

January 20, 2026

Abstract

We present VB, a benchmark for visibility and perspective reasoning in images that explicitly tests whether an AI system can withhold judgement when a human viewer cannot reliably answer. Each example in the benchmark consists of a single photo paired with a short yes/no question about visual evidence, such as whether a sign is readable, whether an object is in view, or whether a person shown in the image can see an object in the image. The AI system is required to output one of three labels (`VISIBLE`, `NOT_VISIBLE`, `ABSTAIN`) plus a confidence for answered predictions. VB is organised by a taxonomy of visibility factors and uses a family-based 2×2 design over a minimal image edit and a minimal text edit. We evaluate models using confidence-aware accuracy with abstention, minimal-edit flip robustness, and a confidence-ranked selective prediction score, and we include a MULTI_AGENT / SECOND_ORDER slice for second-order perspective judgements grounded in a single photo. On 100 families (300 headline items), `llava-onevision` achieves the best overall composite score (0.573) and GPT-4o is second (0.547). We find that text edits are often handled more reliably than minimal image edits, and that only some models produce confidence scores that usefully rank correctness for selective answering.

1 Introduction

Visibility is a prerequisite for safe and reliable image-grounded reasoning. Many visually phrased questions are not answerable from a single photo because the relevant evidence is occluded, out of frame, too small, too dark, washed out by glare, or not visually observable at all. In these cases, systems that guess can appear capable while failing exactly where calibrated withholding of judgement matters.

This paper introduces VB, a benchmark designed to quantify whether a system can:

- verify simple visibility claims from one image and a short question,
- respond appropriately to minimal edits that should flip the correct label,
- abstain when a human viewer cannot reliably and confidently answer from the photo.

We additionally include a MULTI_AGENT / SECOND_ORDER slice that tests second-order perspective judgements (for example, what one person can infer about what another person can see) from a single photo.

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

https://github.com/your-org/vb-benchmark

(Replace with the final repository URL.)

*Affiliation and email.

2 Qualitative examples

Figure 1 illustrates the 2×2 family structure on two concrete families from the dataset. Each family is built from two images (I^0 and I^1) and two questions (q^0 and q^1). In the strict headline subset, the three scored cells follow the XOR pattern described in Section 4.5. The DOUBLE_FLIP cell is retained for diagnostics.

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 judgement. 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, due to blur, occlusion, missing context, or out-of-frame evidence. VizWiz collects questions from blind users, including many with poor image quality or missing evidence [7]. Recent work explicitly evaluates abstention on unanswerable visual questions, for example UNK-VQA [6] and TUBench [8]. VB focuses specifically on visibility and perspective, and uses minimal edits to test whether models change their judgements when evidence changes.

Visibility factors: gaze, occlusion, and field of view. Datasets such as GazeFollow [14] and Following Gaze Across Views [15] study gaze targets and off-frame gaze. Work on occlusion and visibility cues analyses how blockers and partial views affect recognition and localisation. VB differs by evaluating label-level support for a visibility claim under controlled edits rather than requiring precise localisation.

Hallucination and faithfulness in vision-language models. Benchmarks such as MME [4] and POPE [10] 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 judgement when it is not.

Selective prediction, reject options, and risk-coverage. Selective classification formalises 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, 19]. 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 judgements and their brittleness under structured perturbations [9, 18]. 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.

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:

- `VISIBLE`: the claim is supported by visible evidence, defined as "a careful human would answer *yes* from this photo with reasonable confidence";
- `NOT_VISIBLE`: 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?"

Clarifying `NOT_VISIBLE` on negated claims. Because labels refer to the truth of the claim, a question of the form "Is X *not* visible?" can receive gold label `NOT_VISIBLE` when X is plainly visible. In that case, the claim is false and the correct label is `NOT_VISIBLE`, even though X itself is visible.

4.2 Structured outputs

Systems output a single structured prediction:

- `label` $\in$ {`VISIBLE`, `NOT_VISIBLE`, `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`=`VISIBLE`, set `reason_code`=`NONE`. If `label`=`ABSTAIN`, choose exactly one limiting-factor code. If `label`=`NOT_VISIBLE`, 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_VIS

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.
- **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:

$$y^{00} = \texttt{NOT\_VISIBLE}, \quad y^{01} = \texttt{VISIBLE}, \quad y^{10} = \texttt{VISIBLE}, \quad y^{11} = \texttt{NOT\_VISIBLE}.$$

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

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?"	16
LIGHTING_DISTANCE	Darkness, glare, or distance limits	"Can you read the small text on the laptop?"	13
INHERENTLY_NONVISUAL	Properties vision cannot reveal	"Is the device PIN visible anywhere?"	10
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 judgements grounded in one photo	"Does Bob know Alice cannot see the card?"	7
TOTAL			100

4.7 Multi-agent and second-order visibility

The MULTI_AGENT / SECOND_ORDER subset targets second-order judgements, 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 `NOT_VISIBLE`
- `TEXT_FLIP`: (I^0, q^1) with gold `VISIBLE`
- `IMAGE_FLIP`: (I^1, q^0) with gold `VISIBLE`
- `DOUBLE_FLIP`: (I^1, q^1) with gold `NOT_VISIBLE` (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 $y_i \in \{\texttt{VISIBLE}, \texttt{NOT\_VISIBLE}\}$. A model outputs $\hat{y}_i \in \{\texttt{VISIBLE}, \texttt{NOT\_VISIBLE}, \texttt{ABSTAIN}\}$ 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 $\alpha \in [0, 1]$:

$$\text{score}_i = \begin{cases} \alpha & \text{if } \hat{y}_i = \texttt{ABSTAIN}, \\ \hat{c}_i & \text{if } \hat{y}_i = y_i, \\ 1 - \hat{c}_i & \text{if } \hat{y}_i \neq y_i \text{ and } \hat{y}_i \neq \texttt{ABSTAIN}. \end{cases}$$

Then

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

Justification. CAA rewards high confidence when correct and penalises high confidence when incorrect, while giving a fixed partial credit for abstention. It 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 judgement 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).

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

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

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 (AURC)

We evaluate selective prediction using confidence-ranked answering on headline cells. We consider only *answered* predictions (`VISIBLE` or `NOT_VISIBLE`) 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 used in selective classification [3, 19], since $\mathrm{risk}(k) = 1 - \mathrm{acc}(k)$.

We report a normalised score (clamped to $[0, 1]$ for readability):

$$\mathrm{AURC} = \mathrm{clip}_{[0,1]}\left(\frac{A_{\text{model}} - p}{1 - p}\right).$$

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 can occur before clamping when confidence ranking is anti-informative.

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

On the MULTI_AGENT / SECOND_ORDER strict subset, we report:

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

For ToMAcc, `ABSTAIN` counts as incorrect.

5.6 Double flip diagnostic accuracy (DFAcc)

We report diagnostic accuracy on the DOUBLE_FLIP cell:

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

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:

$$\mathrm{FinalScore} = 0.70 \cdot \mathrm{CAA} + 0.15 \cdot \mathrm{MEFR} + 0.10 \cdot \mathrm{AURC} + 0.05 \cdot \mathrm{ToMAcc}.$$

Weight choice. These coefficients reflect a prioritisation: 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 renormalised 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":"VISIBLE|NOT_VISIBLE|ABSTAIN","reason_code":"GAZE_DIRECTION|OCCLUSION|OUT_OF_FRAME|LIGHTING_DISTANCE

Label meanings
```

```
VISIBLE: the claim in the question is clearly true from pixels.

NOT_VISIBLE: 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="VISIBLE", set reason_code="NONE".

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

If label="NOT_VISIBLE":

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 six vision-language models on VB:

- GPT-4o (https://openai.com/index/gpt-4o-system-card/) [13]
- `gemini` (Gemini 1.5 family; https://arxiv.org/abs/2403.05530) [16]
- `llava-onevision` (LLaVA-OneVision; https://arxiv.org/abs/2408.03326) [11]
- `qwen-vl-4bit` (Qwen2-VL; https://arxiv.org/abs/2409.12191) [17]
- `llama-vision` (Llama 3.2 Vision; https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md) [12]
- `claude` (Anthropic Claude; https://platform.claude.com/docs/en/about-claude/models/overview) [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. AURC was computed on answered items only, with `ABSTAIN` excluded from coverage.

All runs produced valid, parseable JSON for all evaluated items (unparsable count 0 for each model).

Table 2: Headline results on VB (300 scored headline items per model). CAA uses $\alpha = 0.25$. AURC is normalised answered-only and clamped to $[0, 1]$. Unclamped normalised AURC is negative for GPT-4o (-0.135), `gemini` (-0.131), and `claude` (-0.048). `ABSTAIN` is the number of abstentions among the 300 headline items.

Model	Final	CAA	I_MEFR	T_MEFR	MEFR	AURC	ToMAcc	ABSTAIN
GPT-4o	0.547	0.635	0.255	0.755	0.505	0.000	0.524	3
llava-onevision	0.573	0.631	0.226	0.925	0.575	0.285	0.333	34
qwen-vl-4bit	0.437	0.468	0.545	0.455	0.500	0.245	0.190	160
claude	0.419	0.504	0.231	0.423	0.327	0.000	0.333	34
gemini	0.378	0.464	0.258	0.194	0.226	0.000	0.381	72
llama-vision	0.353	0.420	0.120	0.400	0.260	0.064	0.286	41

8 Results

8.1 Overall performance

Table 2 reports the main results on the strict headline subset. `llava-onevision` achieves the highest overall FinalScore (0.573), followed by GPT-4o (0.547). `qwen-vl-4bit` (0.437) and `claude` (0.419) form a middle tier, while `gemini` (0.378) and `llama-vision` (0.353) score lower under this weighting.

Abstention behaviour varies widely. GPT-4o abstains rarely (3 of 300 headline items), while `qwen-vl-4bit` abstains frequently (160 of 300). Because CAA awards partial credit for abstention, we report abstention counts alongside all headline metrics.

8.2 Minimal edit sensitivity

Across models, TEXT_FLIP performance often exceeds IMAGE_FLIP performance, suggesting that many failures come from visual sensitivity rather than parsing the text edit. This asymmetry is strongest for `llava-onevision`, which attains high T_MEFR (0.925) but much lower I_MEFR (0.226). GPT-4o shows a similar pattern (0.755 versus 0.255). By contrast, `qwen-vl-4bit` shows more balanced flip rates (I_MEFR 0.545, T_MEFR 0.455), but with lower CAA and substantially higher abstention.

8.3 Abstention, coverage, and confidence ranking

Since AURC 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 unclamped normalised AURC.

Two patterns stand out. First, `qwen-vl-4bit` answers only 46.7% of headline items but achieves the highest answered-only accuracy (0.736), indicating that its abstentions concentrate coverage on easier cases. Second, AURC differs sharply across models: `llava-onevision` and `qwen-vl-4bit` have positive AURC, suggesting informative confidence ranking among answered items, while GPT-4o, `gemini`, and `claude` have negative raw AURC (and are clamped to 0 in Table 2), meaning that confidence ordering is worse than the flat answered-accuracy baseline under this definition.

Table 3: Selective prediction and MEFR diagnostics on headline items. **Answered** counts non-`ABSTAIN` outputs among 300 headline items. **Coverage** is answered/300. **AnsweredAcc** is accuracy on answered items only. $\mathbf{AURC_{raw}}$ is the normalised answered-only score before clamping. **MEFR denom** is the number of families with correct BASE prediction, which determines the denominator in I_MEFR and T_MEFR.

Model	Answered	Coverage (%)	AnsweredAcc	$\mathbf{AURC_{raw}}$	MEFR denom
GPT-4o	297	99.0	0.663	−0.135	94
llava-onevision	266	88.7	0.692	0.285	53
qwen-vl-4bit	140	46.7	0.736	0.245	11
claude	266	88.7	0.545	−0.048	26
gemini	228	76.0	0.553	−0.131	31
llama-vision	259	86.3	0.436	0.064	25

8.4 Multi-agent and second-order subset

ToMAcc is computed on the MULTI_AGENT / SECOND_ORDER slice (7 families, hence 21 headline items). GPT-4o performs best on this subset (0.524). `gemini` is next (0.381). `llava-onevision` and `claude` score 0.333, `llama-vision` scores 0.286, and `qwen-vl-4bit` scores 0.190. The ToMAcc ordering differs from the overall ordering, suggesting that second-order perspective reasoning remains a distinct failure mode even when first-order visibility verification is competitive.

8.5 Diagnostic: double flip

We also record DOUBLE_FLIP behaviour (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.

9 Discussion

VB is designed to make visibility, answerability, and abstention measurable under controlled perturbations. The 2×2 family structure probes two complementary sensitivities: changing the text claim while holding the image fixed, and changing the image evidence while holding the question fixed. The headline score focuses on these three cells, while the double flip is retained for diagnostic analysis of compositional effects. The MULTI_AGENT / SECOND_ORDER slice further isolates second-order perspective judgements that can fail even when first-order verification is straightforward.

The results highlight three recurring themes. First, several models exhibit stronger robustness to text edits than to minimal image edits. This indicates that even when a model follows the question, it may fail to track small, local visual changes that should deterministically flip the correct label. Second, abstention policies differ sharply. A conservative model can abstain frequently, reducing incorrect answers but also reducing coverage, while an aggressive model can answer nearly everything but risk overconfident errors in borderline cases. Reporting both headline CAA and answered coverage is therefore necessary to interpret performance.

Third, confidence values are not consistently useful for selective prediction on this benchmark. Some models show positive AURC, indicating that higher confidence aligns with correctness among

answered items, but others exhibit negative raw AURC, indicating confidence rankings that are anti-informative under this definition. This motivates treating confidence as an empirical interface that must be evaluated, rather than assumed to be meaningful.

Limitations and scope. The current release is modest in scale and collected in a narrow set of environments (campus and nearby streets), so results may not fully represent performance in other settings. The benchmark uses short yes/no questions to tightly control claims, which improves interpretability but does not cover longer multi-step reasoning. Finally, the composite score uses fixed weights to reflect one set of priorities; we emphasise component metrics to support alternative weighting choices.

Implications. Visibility reasoning is a prerequisite for safe image-grounded behaviour. The observed gaps between text-flip and image-flip robustness suggest that improving sensitivity to minimal visual evidence changes remains important. The variability in abstention and in confidence ranking indicates that systems should be evaluated not only on correctness, but also on when they choose to answer and how they express uncertainty.

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. In an evaluation of six vision-language models across 100 families, we find that text-flip robustness often exceeds image-flip robustness, suggesting that brittle visual sensitivity to small evidence changes remains a key failure mode. We also find substantial variation in abstention policies and in whether confidence scores meaningfully rank correctness, underscoring the importance of evaluating selective answering behaviour directly. VB provides a focused stress test for evidence-grounded image understanding and calibrated withholding of judgement from a single photo.

Acknowledgements

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

References

[1] Anthropic. Claude models overview. https://platform.claude.com/docs/en/about-claude/models/overview, 2025. Accessed 2026-01-03.

[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] 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.

[10] 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.

[11] Haotian Liu et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*, arXiv:2408.03326, 2024. URL https://arxiv.org/abs/2408.03326.

[12] Meta. Llama 3.2 vision model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md, 2024. Accessed 2026-01-03.

[13] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. Accessed 2026-01-03.

[14] 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.

[15] 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.

[16] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*, arXiv:2403.05530, 2024. URL https://arxiv.org/abs/2403.05530.

[17] Qwen Team. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint*, arXiv:2409.12191, 2024. URL https://arxiv.org/abs/2409.12191.

[18] 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.

[19] 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.

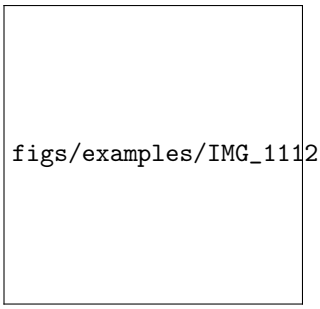
Figure 1: Two example families showing the 2×2 construction. File paths below assume you copy the corresponding images into `figs/examples/`.

Family AV-04 (distance/readability).

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

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

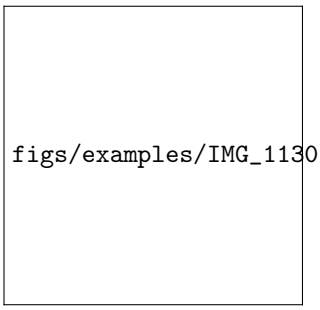
(c) **IMAGE_FLIP** (I^1, q^0).
Gold: `VISIBLE`.



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

Family OC-20 (occlusion).

(e) **BASE** (I^0, q^0).
Q: "Is the metal key blade visible?"
Gold: `NOT_VISIBLE`.



(f) **TEXT_FLIP** (I^0, q^1).
Q: "Is the metal key blade not visible?"
Gold: `VISIBLE`.

(g) **IMAGE_FLIP** (I^1, q^0).
Gold: `VISIBLE`.

(h) **DOUBLE_FLIP** (I^1, q^1).
Gold: `NOT_VISIBLE`.