

# Appendix for PinpointQA

Due to space constraints in the main paper, this appendix supplies supplementary details and analyses. We begin with additional information on dataset construction and a more in-depth description of the evaluation metrics. We then examine failure cases and provide further details on the Human Assistance Evaluation. The appendix concludes with a brief discussion of the benchmark’s scope, limitations, and intended use.

## A. Details of Dataset Construction

1. Scene Curation
2. Reference Objects and Supporting Surface
3. QA Generation
4. Quality Control

## B. Details of Evaluation Metrics

1. FSD Evaluation Setting
2. SSP Evaluation Setting

## C. Representative Qualitative Examples

1. TPV Failure Case
2. NRI Failure Case
3. FSD Failure Case
4. SSP Failure Case

## D. Details of Human Assistance Evaluation

1. Interface and Protocol
2. Scoring Scheme

## E. Discussion

1. Intended Use
2. Limitations

## A Details of Dataset Construction

### A.1 Scene Curation

Scene curation converts each indoor 3D scene into intermediate spatial representations that can be reused across all four tasks. Starting from a predefined small object vocabulary, we identify candidate targets and retain nearby objects as potential references. Instead of using the full scene directly for QA construction, we organize a local neighborhood around each target and preserve only the information most relevant to localization. Nearby objects within 1.0 m are kept by default, and fine-grained 3D spatial relations are computed between the target and its neighbors, including distance and vertical relations. The resulting intermediate spatial representation therefore captures both the nearby objects around the target and their target-centered geometric arrangement, providing a shared grounding basis for subsequent QA generation.

### A.2 Reference Objects and Supporting Surface

Within each intermediate spatial representation, nearby objects play different roles. Reference objects provide local cues that help disambiguate the target position, while the supporting surface serves as the main spatial anchor describing where the target finally rests. This distinction is important because the same grounded context

is reused differently across tasks. In NRI, the supporting surface is excluded so that the question focuses on the nearest valid reference object beyond the most obvious anchor. In FSD, the supporting surface is typically used as the first anchor in the answer, after which one or two nearby reference objects provide additional localization cues. In SSP, the same grounded information is retained in structured form by explicitly preserving the target, its supporting surface, and a small number of nearby references. Supporting surfaces and reference objects are thus complementary rather than interchangeable.

### Small Object Vocabulary

bag, backpack, book, bottle, bowl, candle, cup, dumbbell, hair dryer, hat, headphones, keyboard, laptop, mouse, purse, shoe, tissue box, water bottle, alarm clock, ball, stuffed animal, plate, clock, cd case, coffee kettle, paper towel roll, scale, soap dish, soap dispenser, telephone, toaster, tray, water pitcher, fire extinguisher, paper cutter, plunger, dustpan, broom, power strip, toilet paper, guitar, toaster oven, paper bag, key, phone, smartphone, mobile phone, earbuds, earphones, wireless headphones, headset, headphone case, remote control, tv remote, remote controller, remote, tv controller, game controller, charger, power adapter, laptop charger, phone charger, power bank, wireless charger, power plug, adapter, dongle, usb hub, extension cord, power cord, cable, wallet, handbag, tote bag, glasses, eyeglasses, glasses case, mouse pad, mousepad, pen, pencil, marker, highlighter, notebook, notepad, mug, spray bottle, lotion bottle, deodorant bottle, toothbrush, razor, comb, hairbrush, scissors, stapler, tape dispenser, glue bottle, calculator, lanyard, pen holder, pencil holder, pencil case, post it note, sticky note, phone stand, tissue, eraser, soap, power extension

### A.3 QA Generation

Based on the intermediate spatial representation, we instantiate four task formats of increasing difficulty. For TPV, we generate both positive and negative presence queries; negative queries are drawn from small-object categories absent from the scene, so both labels remain grounded in scene-level object presence. For NRI, we construct multiple-choice questions over valid reference objects and ask the model to identify the nearest one while excluding the supporting surface. For FSD, we transform the same grounded spatial facts into concise but precise free-form answers, usually starting from the supporting surface and then incorporating one or

two nearby reference objects with centimeter-level distances. For SSP, we preserve the same localization-relevant information in a parsable JSON format containing the target, the supporting surface, and a small set of nearby references. These four tasks are therefore not built from separate annotation pipelines, but from different instantiations of the same intermediate spatial representation.

## A.4 Quality Control

To improve dataset reliability, we combine automatic filtering with iterative manual review. On the automatic side, we remove invalid labels and background or structural objects during scene curation, and we generate QA pairs only for target instances satisfying task-specific validity constraints. In particular, multiple-choice questions must contain four distinct options, free-form answers must remain human-readable, and structured outputs must be directly parsable. Samples or scenes that fail these checks are excluded rather than retained as noisy supervision. On the manual side, each review round spot-checks roughly 100 QA pairs spanning about 10–15 scenes. We examine whether Task 2 mistakenly includes the supporting surface as a candidate, whether Task 3 correctly preserves the supporting surface and nearby references, whether distances and relations remain consistent with the underlying geometry, whether the templates are natural and well formed, and whether the resulting localization is plausible from a human observation perspective. Issues found in these checks are then used to revise QA generation logic, question templates, and filtering rules, while keeping the underlying intermediate spatial representation unchanged.

## B Details of Evaluation Metrics

### B.1 FSD Evaluation Setting

Here we provide additional implementation details for the GPT-5.4-based judge used in FSD evaluation. To improve scoring stability, the judge is required to return a structured JSON object rather than free-form commentary. The output includes the total score, the five dimension scores defined in the main paper, and explicit error tags such as wrong main location, wrong supporting surface, missing key reference, wrong spatial relation, missing or wrong numeric distance, unclear expression, and hallucinated reference.

In our implementation, the returned total score is also checked against the sum of the five dimension scores and automatically corrected if the two disagree, which avoids accidental inconsistencies in judge outputs. This design makes the metric not only scalar, but also diagnostic. For completeness, we provide the scoring rubric in Figure 1, while the full judge prompt is included in the project repository.

### B.2 SSP Evaluation Setting

Here we provide additional implementation details for the SSP metric described in the main paper. Before scoring, the evaluator first attempts to recover a valid JSON object from the model output, even when the prediction is wrapped in extra text or code fences. It then applies conservative normalization to object labels, supporting-surface labels, and relation tokens, so that superficial formatting differences do not dominate the score. The recovered output is also checked for basic structural validity, including the presence of a

supporting surface field and a references list with object, relation, and distance entries.

After normalization, predicted references are aligned to the ground-truth slots before computing the weighted soft score defined in the main text. Exact relation matches receive full credit, while the limited set of semantically close relation pairs listed in the main paper receive partial credit. Distance is scored using the same range-aware tolerance policy as FSD. In this way, the metric remains tolerant to formatting noise while preserving strict evaluation of localization-relevant structured content.

## C Representative Qualitative Examples

This section presents representative qualitative examples to analyze the typical failure modes of different models in small-object spatial understanding. We examine three scenes with different small-object targets: *mouse* in scene 062e5a23a6, *mobile phone* in scene 0b031f3119, and *headphone case* in scene 281ba69af1. As shown in Figures 2, 3, and 4, the errors do not arise in a single uniform way. Instead, different models break down at different stages, ranging from missing the target itself, to selecting the wrong nearby anchor, to losing the correct target-centered local spatial context in free-form description and structured prediction. In all figures, green marks indicate GT-consistent content, while red marks indicate content that deviates from the GT.

### C.1 TPV Failure Case

We first examine failures at the Target Presence Verification (TPV) stage. Figure 2 shows a representative TPV error for *mouse* in scene 062e5a23a6. Although the target is present in the scene, Cambrian-S-7B directly predicts *No*. This case reflects one of the earliest failures in the progressive capability chain: even when the overall desktop scene is clearly visible, the model still cannot reliably determine whether the queried target is present. More importantly, it suggests that, in cluttered indoor videos, perceiving the overall scene does not necessarily translate into accurate judgments about the presence of a small target. If the model cannot complete target presence verification at the outset, then later stages, including Nearest Reference Identification, Fine-Grained Spatial Description, and Structured Spatial Prediction, also lose their reliable foundation. This also indicates that current general-purpose MLLMs remain weak at preserving small-object-related cues, especially when the target itself is not salient and is surrounded by more visually prominent distractors.

### C.2 NRI Failure Case

At the Nearest Reference Identification (NRI) stage, the same *mouse* example also shows a typical error. As shown in Figure 2, although GPT-5.4 is able to determine that the target appears in the scene, it still makes an error when selecting the nearest reference object: instead of choosing the GT answer *office chair*, it predicts *laptop bag*. Compared with TPV, NRI requires the model not only to recognize the queried target, but also to determine which nearby object provides the nearest and most informative cue for localization. This example shows that, even after recognizing the target, the model may still fail to identify the most important nearby reference around it. Instead, it is more easily distracted by objects that

are visually more prominent in the scene, and therefore selects an object that looks plausible but is not actually the nearest one in spatial terms. As a result, the transition from Target Presence Verification to Nearest Reference Identification remains a clear challenge for current models. More broadly, this also suggests that general visual-language models often rely on salient objects when forming an answer, rather than consistently reasoning around the target’s nearby local spatial context.

### C.3 FSD Failure Case

The Fine-Grained Spatial Description (FSD) examples in Figures 2, 3, and 4 show a fairly consistent pattern: the generated descriptions are often fluent and seem reasonable at the overall scene level, but the local anchors used for localization have already drifted away from the GT. In Figure 2, for example, LLaVA-OneVision-1.5-8B describes the *mouse* using *laptop* and *keyboard*, whereas the truly important reference objects in the GT are *office chair* and *glasses case*. In Figure 3, Kimi K2.5 rewrites the *mobile phone* into a more generic desk-level description centered on *monitor* and *keyboard*, rather than preserving the GT description built around *tripod*, *table*, and *keyboard*; InternVL3.5-8B-Instruct retains part of the tabletop context, but still shifts the local configuration to *keyboard* and *mouse*, with distances that no longer match the GT. In Figure 4, GPT-5.4 preserves the target category and the broad tabletop context, but its local references drift to *cloth*, *tissue packet*, and *plant*, while LLaVA-OneVision-1.5-8B goes further and places the target entirely on the *floor*. Taken together, these examples show that the difficulty of FSD does not lie in generating natural language itself, but in preserving the correct target-centered local spatial context while turning perception into text. In other words, a description that sounds natural does not necessarily remain spatially accurate. More broadly, this also suggests that once the model enters free-form generation, its output is easily shaped by common expressions and scene priors; if the local spatial evidence around the target is not stably retained, the model is more likely to fall back on descriptions that are more familiar and more fluent, but already deviate from the GT in spatial terms.

### C.4 SSP Failure Case

A similar pattern can also be observed at the Structured Spatial Prediction (SSP) stage, except that the errors now appear in structured outputs. Figure 2 shows that SenseNova-SI-1.3-InternVL3-8B generates a syntactically valid JSON output for *mouse*, but the support surface, reference objects, and spatial relations no longer match the GT. Figures 3 and 4 further illustrate the same issue. For *mobile phone*, Kimi K2.5 produces a well-formed JSON object, but replaces the original GT support surface and reference structure with a desk-centered configuration involving *keyboard* and *monitor*; InternVL3.5-8B-Instruct retains *table* as the support surface, but still shifts the nearby reference objects and their spatial relations to *keyboard* and *mouse*, thereby breaking the original target-centered spatial representation. For *headphone case*, GPT-5.4 outputs a valid JSON object and preserves the broad tabletop context, but its reference objects drift to *tissue box* and *door*; SenseNova-SI-1.3-InternVL3-8B further rewrites the support surface as *chair*, causing the local configuration around the target to deviate even

further from the GT. Taken together, these examples show that SSP is not merely a test of whether a model can output a well-formed JSON object. More importantly, it examines whether the model can preserve the target-centered spatial evidence established in earlier stages within a structured representation. From the model outputs, a clear gap can be seen: structural correctness constrains the format of the output, but does not guarantee that the spatial representation itself remains correct. Once the earlier visual localization has already drifted, the model may still produce a fully valid JSON object whose spatial content is nevertheless wrong.

Overall, these three examples reflect an error pattern that becomes increasingly apparent as the task progresses. Some models fail at relatively early stages, either by failing to determine whether the target is present or by failing to identify the nearest reference object correctly. Others are able to pass the earlier stages, but gradually lose the correct local anchors for localization when generating free-form descriptions or structured outputs. This suggests that the central challenge highlighted by PinpointQA is not simply whether a model can mention the target object, but whether it can maintain an accurate and consistent target-centered spatial representation throughout the full prediction chain.

## D Details of Human Assistance Evaluation

### D.1 Interface and Protocol

We implement the Human Assistance Evaluation using a web-based click interface. For each sample, the interface displays the question together with multiple video frames. Participants browse the frames, select the one most likely to contain the correct target location, and click the estimated target location. The visual interaction is identical across all settings, and only the accompanying text differs. In the Unguided setting, participants see only the question and the video frames. In the two guided settings, the textual hint is shown before the participant begins answering the question. In the Model-Assisted setting, participants additionally see the natural-language FSD answer generated by the fine-tuned Qwen3-VL-8B-Instruct-SFT model. In the GT-Assisted setting, this text is replaced by the ground-truth FSD answer. Figure 5 shows the interface under the assisted setting.

For each sample, the participant starts the question, browses the frames, clicks the target location on the selected frame, and saves the response. Ground-truth target points are annotated beforehand using the same tool in annotation mode, which records the reference frame index and normalized click coordinates for each sample.

### D.2 Scoring Scheme

Each submitted response consists of a selected frame index and normalized click coordinates  $(x, y)$ . The completion time for each question is measured from the moment the participant starts the question to the moment the response is saved. Accuracy is evaluated against the manually annotated target point. If the selected frame is incorrect, the score is set to 0. Otherwise, let  $d$  denote the normalized Euclidean distance between the participant’s click and the annotated target point. The score is computed with a quadratic

decay within a fixed radius  $r = 0.12$ :

$$s = \max\left(0, 1 - \left(\frac{d}{r}\right)^2\right), \quad r = 0.12.$$

The final accuracy is reported as the mean of per-question scores, and the completion time is reported as the mean of per-question elapsed times.

This metric requires both correct frame selection and accurate localization within the chosen frame. Compared with frame-only retrieval accuracy, it imposes a stricter criterion while still providing a graded measure rather than a binary decision.

## E Discussion

### E.1 Intended Use

PinpointQA is intended as a focused benchmark for small object-centric spatial understanding in indoor videos. Rather than covering all aspects of video understanding, it is designed to probe a specific capability chain: from target presence, to reference-based grounding, to fine-grained spatial description, and finally to structured spatial output. This makes it particularly suitable for studying how multimodal systems localize small objects through nearby references and how well such grounded information can be expressed in both human-readable and machine-usable forms. In addition to benchmarking, the experimental results also suggest that PinpointQA can serve as a useful source of supervision for instruction tuning or supervised fine-tuning on grounded spatial QA tasks, especially when the goal is to improve localization-oriented behavior rather than broad open-domain video understanding.

### E.2 Limitations

At the same time, the scope of PinpointQA is intentionally restricted, and this boundary should be considered when interpreting results. The benchmark is limited to indoor scenes and to small object-centric localization and spatial expression, rather than full-scene understanding, open-world tracking, dense video captioning, or unconstrained embodied interaction. Moreover, the released QA pairs are derived from intermediate spatial representations and benchmark construction logic, which improves grounding consistency and comparability across tasks, but can also make some answer styles more regular than unconstrained human language. This reflects a deliberate trade-off in the benchmark design: PinpointQA prioritizes stable, localization-relevant supervision over maximum linguistic or environmental diversity. In addition, because the repository does not redistribute original videos or raw scene assets, full reproduction still depends on separate access to the underlying source datasets.

## Scoring Rubric

Score the answer on five dimensions, for a total of 10 points.

### A. Main location and support surface (0–3)

Judge whether the answer identifies the correct main location, primary support surface, or anchor region.

3: Main location and support surface are correct, and the overall placement matches the Ground Truth.

2: The main support surface or anchor is correct, but the placement is still incomplete because one or more important localizing details are missing.

1: The answer only identifies a broad related area or generic region, without a reliable support surface, anchor, or precise local placement.

0: Main location or support surface is clearly wrong.

If the object is placed on a clearly wrong support surface or in a clearly wrong scene region, this dimension must be 0 or 1. If the predicted main location is unrelated to or opposite from the Ground Truth, it should be 0.

### B. Key reference objects (0–2)

Judge whether the answer mentions the correct nearby reference objects from the Ground Truth.

2: At least two correct and clearly relevant key references are mentioned, and they play a role in localization.

1: Exactly one correct and clearly localizing Ground-Truth-supported reference is mentioned and used meaningfully for localization.

0: No clearly localizing Ground-Truth-supported reference is provided, or the references are vague, generic, mostly wrong, unsupported, or merely co-mentioned without helping localization.

Only Ground-Truth-supported references receive credit.

### C. Spatial relations (0–2)

Judge whether the answer correctly states the relation between the target and the key references.

2: The important relations are basically correct.

1: At least one important directional relation remains substantially correct, but another relation is incomplete, less precise than the Ground Truth, or mixed with some error.

0: Important relations are wrong, missing for most key references, or replaced mainly by vague proximity.

If the Ground Truth uses a strong directional relation and the model weakens it to vague proximity, reduce the score. If the Ground Truth uses a directional relation (e.g., left, right, behind, under, above, below) but the model replaces it with vague proximity terms such as “near” or “close to,” this should usually receive 0 and at most 1 only when another important directional relation is still substantially correct. If a key direction is reversed, penalize heavily.

For this task, vague proximity is not a substitute for directional localization when the Ground Truth provides a directional relation.

### D. Centimeter-level distance (0–2)

Judge whether the answer provides usable numeric distance information in centimeters and whether it is roughly consistent with the Ground Truth.

2: The answer includes usable numeric distance information and it is roughly consistent with the Ground Truth.

1: The answer includes at least one usable numeric distance in centimeters (or directly convertible metric form) tied to a correct Ground-Truth-supported reference, but the value is somewhat inaccurate or incomplete.

0: There is no usable numeric distance, only vague wording, the number lacks a usable unit, the number is not tied to a correct reference, or the numeric information is unreliable.

Treat “8 cm,” “8cm,” “about 8 cm,” and “8 centimeters” as the same type of usable distance. Vague phrases such as “very close” or “a few centimeters” do not count. A bare number without a unit does not count unless the context clearly implies centimeters. Metric expressions directly convertible to centimeters, such as 0.1 m, may be treated as usable. Answers that only provide approximate or generic distance descriptions without specific numeric grounding should receive 0. If the Ground Truth provides multiple key reference distances but the answer gives only one weakly grounded or mismatched number, this should usually receive 0 rather than 1. As rough tolerance, allow about  $\pm 5$  cm for Ground-Truth distances in the 0–10 cm range,  $\pm 10$  cm in the 10–30 cm range, and  $\pm 15$  cm for distances above 30 cm.

### E. Clarity (0–1)

Judge only whether the answer is understandable.

1: The answer is clear and readable.

0: The answer is severely confusing, malformed, or hard to understand.

This dimension has very low weight and must not compensate for localization errors.

Compute  $\text{total\_score} = A + B + C + D + E$ . The total score must be an integer from 0 to 10. If  $A = 0$  and  $B = 0$ , the total score should not exceed 2. If the answer contains almost no usable localization information, it should usually not exceed 2. If it hallucinates a wrong main scene or wrong key references, score strictly.

**Figure 1: Core scoring rubric of our prompt-based judge for Fine-Grained Spatial Description. The full prompt is available in the project repository.**



Scene: 062e5a23a6

**TPV**

Question: In the entire scene, did the **mouse** appear?  
 Ground Truth: **Yes**

**NRI**

Question: Excluding the surface it is resting on, which of the following objects is nearest to the final location of the **mouse** in this scene?

A. glasses case      B. monitor  
 C. office chair      D. laptop bag

Ground Truth: **C. office chair**

**FSD**

Question: Please describe in detail the final location of the **mouse** in the scene, including its distance and relation to nearby objects.

Ground Truth: **The mouse is on the table. It is also to the side of and above the office chair, about 6 cm away. It is further near the glasses case, about 26 cm away.**

**SSP**

Question: Please output the final location of the **mouse** in structured JSON format.

Ground Truth: **{ "target": "mouse", "support\_surface": "table", "references": [ { "object": "office chair", "relation": "side\_above", "distance\_cm": 5.8 }, { "object": "glasses case", "relation": "near", "distance\_cm": 25.6 } ] }**

**TPV (Cambrian-S-7B)**

Question: In the entire scene, did the mouse appear?  
 Output: **No**

**NRI (GPT-5.4)**

Question: Excluding the surface it is resting on, which of the following objects is nearest to the final location of the mouse in this scene?  
 Output: **D. laptop bag**

**FSD (LLaVA-OneVision-1.5-8B)**

Question: Please describe in detail the final location of the mouse in the scene, including its distance and relation to nearby objects.

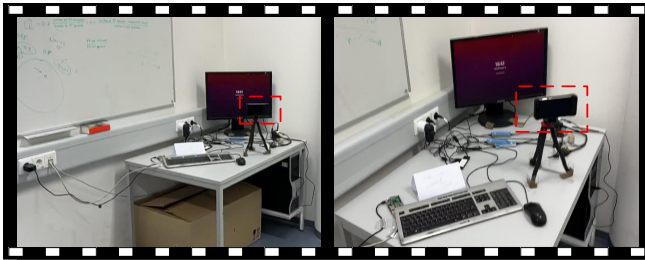
Output: **The mouse is located on the desk, approximately 23 cm away from the edge of the desk. It is positioned near a black laptop, which is about 41 cm away from the mouse. The desk also has a white keyboard placed directly in front of the mouse, at a distance of about 15 cm.**

**SSP (SenseNova-SI-1.3-InternVL3-8B)**

Question: Please output the final location of the mouse in structured JSON format.

Output: **\texttt{\{"target": "mouse", "support\\_surface": "chair", "references": [ \ {"object": "chair", "relation": "on"}, \ {"object": "desk", "relation": "below"} ] \}}**

Figure 2: Example 1. Progressive failure across TPV, NRI, FSD, and SSP for target *mouse* in scene 062e5a23a6. Different models break down at different stages, ranging from missing the target itself, to selecting an incorrect nearby anchor, to producing free-form and structured outputs that no longer preserve the correct target-centered local spatial context.



Scene: 0b031f3119

FSD

Question: Please describe in detail the final location of the **mobile phone** in the scene, including its distance and relation to nearby objects.

Ground Truth: The mobile phone is above the tripod. It is also to the side of and above the table, about 23 cm away. It is further to the side of and above the keyboard, about 27 cm away.

SSP

Question: Please output the final location of the **mobile phone** in structured JSON format.

```
Ground Truth: {"target": "mobile phone",
"support_surface": "table", "references":
[{"object": "tripod", "relation": "above",
"distance_cm": 2.4}, {"object": "table",
"relation": "side_above", "distance_cm": 23.1}]}
```

FSD

Kimi K2.5

Output: The mobile phone rests on the white desk surface, positioned approximately 15 cm to the right of the black computer monitor and about 25 cm in front of the wall-mounted power outlets. The phone lies flat on the desk, separated from the keyboard by a small gap of roughly 20 cm.

InternVL3.5-8B-Instruct

Output: The mobile phone is on the table, 23 cm away from the keyboard and 41 cm from the mouse.

SSP

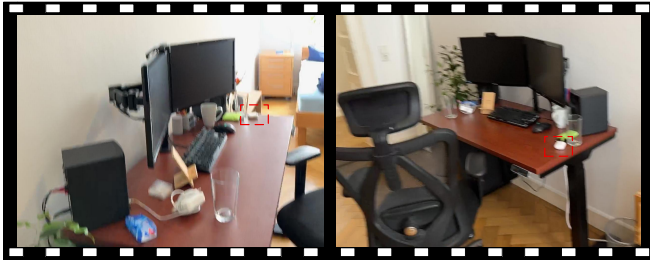
Kimi K2.5

```
{"target": "mobile phone", "support_surface":
"desk", "references": [{"object": "keyboard",
"relation": "next_to", "distance_cm": 15.0},
{"object": "monitor", "relation": "near",
"distance_cm": 25.0}]}
```

InternVL3.5-8B-Instruct

```
{"target": "mobile phone", "support_surface":
"table", "references": [{"object": "keyboard",
"relation": "side_below", "distance_cm":
15.0}, {"object": "mouse", "relation":
"side_below", "distance_cm": 12.0}]}
```

Figure 3: Example 2. Drift of target-centered local spatial context in FSD and SSP for target *mobile phone* in scene 0b031f3119. Both Kimi K2.5 and InternVL3.5-8B-Instruct produce scene-plausible but incorrectly grounded predictions, with errors in support surface, reference objects, spatial relations, and distances.



Scene: 281ba69af1

FSD

Question: Please describe in detail the final location of the **headphone case** in the scene, including its distance and relation to nearby objects.

Ground Truth: The **headphone case** is on the table. It is also to the side of and below the speaker, about 12 cm away. It is further to the side of and below the monitor, about 17 cm away.

SSP

Question: Please output the final location of the **headphone case** in structured JSON format.

Ground Truth: `{"target": "headphone case", "support_surface": "table", "references": [{"object": "speaker", "relation": "side_below", "distance_cm": 12.0}, {"object": "monitor", "relation": "side_below", "distance_cm": 17.1}]}`

FSD

GPT-5.4

Output: The **headphone case** ends up on the wooden desk along the wall, near the left side of the tabletop. It's right beside a folded gray cloth and about 20-30 cm from a white tissue packet, with a plant just off to the desk's right.

LLaVA-OneVision-1.5-8B

Output: The **headphone case** is located on the floor, approximately 23 cm away from the edge of the desk. It is positioned next to a black chair, which is about 41 cm away from the desk. The case appears to be resting on a blue towel that is spread out on the wooden floor.

SSP

GPT-5.4

Output: `{"target": "headphone case", "support_surface": "desk", "references": [{"object": "tissue box", "relation": "next_to", "distance_cm": 10.0}, {"object": "door", "relation": "near", "distance_cm": 20.0}]}`

SenseNova-SI-1.3-InternVL3-8B

Output: `{"target": "headphone case", "support_surface": "chair", "references": [{"object": "chair", "relation": "on", "distance_cm": 0.5}, {"object": "table", "relation": "next_to", "distance_cm": 1.7}]}`

Figure 4: Example 3. Inconsistent grounding across free-form and structured outputs for target *headphone case* in scene 281ba69af1. Even when the target category is preserved, the models fail to maintain the same target-centered local spatial evidence, leading to incorrect support surfaces, reference objects, and spatial relations.

Mode

Annotation  Test

Test Set

Raw  Model  GT

Progress  
11 / 20

Sample Info  
Set: GT  
ID: 20871b98f3  
Images: 4

Timer  
00:13

Session Summary  
GT run ready. Results will be saved after the final sample.

QUESTION

**Where is the scissors? Please click on its location.**


GROUND TRUTH

The scissors is to the side of and above the dish drying rack, about 4 cm away. It is also to the side of and below the neon light, about 9 cm away. It is further to the side of and below the kitchen cabinet, about 12 cm away.

Image 3 / 4 Select a frame, click the target location, then save to continue.

← Previous

2.png



3.png

Next →

4.png

Start Test Save Answer

Answer point selected on image 3. Save to continue.

Figure 5: Human evaluation interface in the GT-Assisted setting.