

# Office Hours: Supplementary Material

Anonymous Author(s)  
Affiliation  
Address  
email

## 1 Dataset Statistics

The dataset composition and question distribution are summarized in Table 1. Furthermore, for each cubicle in video 0 of local changes and global changes subsets, we provide a manually curated list of objects initially present in the scene. This enables researchers to identify the full set of objects in each video using both the object list and the corresponding changes datasheet. We provide examples of the local and global object lists in Table 2. In total, the initial scenes include 163 distinct objects in the local changes subset and 541 distinct objects in the global changes subset.

Table 1: Dataset Statistics

Questions Statistics						
Category / Person	Counting	Detection	Location	State	Cubicle	Total
<b>Global Questions</b>	126	115	123	126	–	490
<b>Static Association–Semantic Mapping VQA</b>	29	43	51	36	30	189
<b>Local Questions</b>						
Alana 2038Q	25	25	25	25	–	100
Alexandria 2008M	25	26	25	25	–	101
Amin 2008E	25	26	25	23	–	99
Daniel 2038Y	26	24	24	22	–	96
Emily 2038U	25	25	25	24	–	99
Fernando 2041S	24	25	25	25	–	99
Jack S. 2038T	25	27	24	25	–	101
Jason 2008K	26	26	26	27	–	105
Josh 2008S	23	24	24	24	–	95
Kristine 2038N	25	25	24	25	–	99
<b>Local Total</b>	249	253	247	245	–	994
<b>Grand Total</b>	<b>376</b>	<b>362</b>	<b>373</b>	<b>371</b>	<b>30</b>	<b>1671</b>
Video Statistics						
<b>Number of videos</b>						221
<b>Minutes of Footage</b>						136

Table 2: **Initial objects in two cubicles.** Local = 2038U (“Emily”); Global = 2038S (“Frank”).

Local (2038U)		Global (2038S)	
Object	Count	Object	Count
Computer Monitor	2	chair	1
Computer	1	whiteboard	1
Keyboard	1	laptop	1
Office Chair	1	stack of paper	2
Laptop	2	keyboard	1
Book	1	monitor	1
GEFORCE RTX 4090 cardboard box	1	monitor stand	1
Whiteboard	1	red, black medium rectangular box	1
Marker	5	white paper	3
Glass jar	1	medium black box	2
Hat	1	mouse	1
Extension outlet	1	white cardboard box	1
Mouse	1	plastic clear long thin object	1
Monitor stand	2	NVIDIA geforce box	1
3D printed cat figurine	1	black and yellow small object	1
3D printed ship figurine	1	pen	2
3D printed buddha figurine	1	outlet surge protector	1

## 2 Prompts

This section, we provide the prompts that were feed into Gemini (2.5 Flash [1] & 2.5 Pro [2]) for generating visual question answering (VQA) questions, along with the prompts for benchmark evaluation.

### 2.1 Prompt for Static Association Semantic Mapping VQA Creation

The Static Association–Semantic Mapping VQA benchmark measures the VLM’s ability to link static visual cues from a single global video to the spatial information encoded in a semantic map.

Questions are generated by supplying Gemini 2.5 Pro [2] with the keyframes (extracted as shown in Section 5) and the prompt listed in Listing 1. In contrast to the Local and Global Video-Change VQA sets, this benchmark introduces a fifth category—Cubicle/Room Location—which targets spatial reasoning in static scenes. An example of this new category, produced by Gemini, is shown in Table 4.

```

You are a highly accurate visual reasoning assistant.

Your task is to generate exactly one question per category based strictly on
→ the visible content of the image and the provided context. Use natural and
→ casual language that sounds like something a person would ask when viewing the
→ image.

The categories are:
- Object detection / classification
- Object counting
- Object state changes (attribute)
- Object location
- Cubicle/room location

Context for the image:
{prefix}

Strict instructions:
- Only use objects, attributes, and relationships that are clearly visible in
→ the image. Do not guess or hallucinate.

```

```

- Use the context (robot location, viewed areas, landmarks) to make the question
  ↳ specific. Viewed areas are based on the robot's perspective (closest to
  ↳ furthest cubicle).
- Refer to cubicles using aliases when available (e.g., "Harish's cubicle"). Don't
  ↳ mention the robot's position in the question.
- Avoid yes/no questions.
- Questions must be focused on objects **inside cubicles**.
- For "Cubicle/room location", use spatial comparisons (e.g., relative to another
  ↳ cubicle or a door).
- Make sure the correct answer is visibly justifiable \- otherwise skip that
  ↳ object/question.
- Randomize the correct answer position (A-D). "E" must always be "None of the
  ↳ above".
- **Do not include questions for any other categories. Return only the five
  ↳ specified.**

Respond with five clearly separated JSON blocks, each prefixed by a header like
↳ this:

[Object detection / classification]
{
  "Type": "Object detection / classification",
  "Image": "{image_id}",
  "Question": "...",
  "Multiple Choice": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "...",
    "E": "None of the above"
  },
  "Correct Choice": "<A|B|C|D|E>"
}

Only use the categories listed. Do not invent new ones or return fewer/more than
↳ five total questions.

```

Listing 1: Prompt used to create the static association semantic mapping VQA Generation.

## 19 2.2 Prompt for Static Association Semantic Mapping VQA Benchmark

20 To evaluate the Static Association Semantic Mapping VQA Benchmark, Gemini 2.5 Pro [2] was given  
 21 one global video in combination with the prompt shown in Listing 2 and all of the static association  
 22 semantic mapping questions and answer choices. This was done using Google AI Studio.

```

Answer the following questions based on the video.

```

Listing 2: Prompt utilized in Static Association Semantic Mapping VQA Benchmark

## 23 2.3 Prompt for Global/Local Video Changes Questions Creation

24 The global and local changes were partitioned by category into four CSV files: *Object Counting*,  
 25 *Object Detection*, *Object Location*, and *Object State*. Each CSV and an accompanying by prompt  
 26 shown in Listing 3 (global) and Listing 4 (local), were supplied to Gemini 2.5 Flash [1], which  
 27 generated one question per change. A random sample of 20 questions per category was subsequently  
 28 validated by a human annotator for correctness and clarity.

29 Every generated question was required to be in a five-option multiple-choice format (A–E) with  
 30 choice E reading “None of the above” (or equivalent), to demand multimodal reasoning—so that the  
 31 correct answer could not be inferred from the text description alone—and to hinge on the temporal  
 32 comparison of two consecutive videos. Examples of the questions created by Gemini are shown in  
 33 the Table 3.

You are given the Object Counting CSV file, which is one of four CSV files that  
→ describe changes between consecutive videos in a multi-video dataset:

CSV file	What it describes
→ Example	
→ -----	-----
→ -----	-----
<b>**Object Location Change**</b>	Where a unique object moved
→ 'Black mug moved from desk A to desk B'	
<b>**Object Detection**</b>	Appearance or disappearance of a unique object
→ 'New coffee machine appeared'	
<b>**Object Counting**</b>	Change in the count of a non-unique object
→ 'Number of blue pens decreased from 5 to 3'	
<b>**Object State Change**</b>	State transition of a unique object
→ 'Monitor changed from <i>*off*</i> to <i>*on*</i> '	

The changes are cumulative meaning we are adding changes from video to video.  
All the videos are the same between the CSV files. Meaning the videos representing  
→ episode 1 are the same across all csv files provided.

#### #### Your task

From the single CSV I've provided, create **\*\*one multiple-choice question for**  
→ ***\*every\** recorded change\*\***.  
These questions will be used to benchmark a Vision-Language Model (VLM).

---

#### #### Question requirements

Note: word inside [] mean variables.

1. The question must follow one of the following formats:
  - \* In [cubicle's] cubicle, what object's quantity has been changed?
    - Answer must include [object] and the [initial-count] and  
→ [final-count]
  - \* In [cubicle's] cubicle, how did the quantity of [object] changed?
    - Answer must include the [initial-count] and [final-count]
2. Cover every change in the CSV once and only once.
5. No video numbers in the question text.
6. Choices: exactly five (A-E).
  - \* Choice E must be a catch-all such as 'None of the above' or 'No change' or  
→ something similar.
  - \* From time to time the correct answer must be E.
7. Correctness: specify the correct letter.

---

#### #### JSON Requirements

1. **\*\*Type\*\*** must be one of:
  - \* **`Object Location Change`**
  - \* **`Object Detection`**
  - \* **`Object Counting`**
  - \* **`Object State Change`**
2. Change number must be present in the JSON output for every question
3. Initial Video and Final video must be present on JSON output with the following  
→ format. Final\_video = Episode and Initial Video = Episode - 1. Therefore if  
→ episode is 1 then initial video is 0 and final video is 1.
4. The question must be present on the JSON format.
5. The multiple choice questions must be present.
6. The correct choice must be present.

#### #### Output format (JSON)

Please follow the following format.

```
```json
{
  "Q1": {
    "Type": "Object State Change",
    "Change Number": 7,
    "Initial Video": 3,
    "Final Video": 4,
    "Question": "What change occurred to the desk lamp?",
    "Multiple Choice": {
      "A": "It moved from the left shelf to the center shelf.",
      "B": "It changed from 'off' to 'on'.",
      "C": "A second lamp appeared next to it.",
      "D": "It was replaced by a white mug.",
      "E": "None of the above."
    },
    "Correct Choice": "B"
  },
  "...": {}
}
```
```

Please do not execute any code to do this. Just read the csv line by line and make  
→ the questions.  
Finally make sure the question has proper english grammar.  
*\*Repeat this structure for every change in the CSV.\**

Listing 3: Example of the prompt feed into Gemini for the creation of the object counting questions of the global video changes.

You have been provided one CSV file named "Object Detection.csv", from a total of  
→ four possible CSV files:

1. Object Location Change.csv
2. Object Detection.csv
3. Object State Change.csv
4. Object Counting.csv

#### ## Purpose and Context

The objective is to generate natural and realistic questions aimed at benchmarking  
→ VLM abilities to understand and identify object associations, appearances, and  
→ disappearances within realistic office cubicle scenes between consecutive  
→ video episodes.

#### ## Explanation of CSV Files

- Each CSV file represents changes observed between consecutive videos.
  - Object Location Change.csv: Tracks movements of objects between locations  
→ within the cubicle.
  - Object Detection.csv: Tracks the appearance or disappearance of unique objects  
→ within the cubicle.
  - Object State Change.csv: Captures state transitions (initial to final state)  
→ of objects.
  - Object Counting.csv: Documents quantity changes of non-unique objects within  
→ the cubicle.
- Changes accumulate progressively across episodes (video n-1 to video n).
- Videos referenced across CSV files for the same episode are identical.

#### ## Explanation of Columns in "Object Detection.csv"

- Episode: Indicates the videos between which changes occur. Episode n is from  
→ video (n-1) to video n.
- Unique Object: Specifies the unique object that appeared or disappeared during  
→ the episode.

- Location inside Cubicle: Indicates the location of the unique object within the  
→ cubicle.
- Appear/ Disappear (Change): Indicates whether the object appeared or disappeared  
→ in the episode.

## ## Requirements for Generating Questions

Generate a JSON file containing questions based on every single recorded change  
→ within the provided CSV file. The JSON structure for each question must be as  
→ follows:

```
```json
{
  "Q1": {
    "Type": "Object Detection",
    "Change Number": 1,
    "Initial Video": 0,
    "Final Video": 1,
    "Question": "The question text",
    "Multiple Choice": {
      "A": "Possible Answer A",
      "B": "Possible Answer B",
      "C": "Possible Answer C",
      "D": "Possible Answer D",
      "E": "None of the above"
    },
    "Correct Choice": "A"
  }
}
```
```

## ## Detailed Instructions

1. Increment each question entry label sequentially (Q1, Q2, etc.).
2. Clearly specify "Type" from: Object Location Change, Object Detection, Object  
→ Counting, Object State Change. For the current CSV provided ("Object  
→ Detection.csv"), type is always "Object Detection".
3. "Change Number": Clearly corresponds to each row entry in the CSV for easy  
→ reference.
4. "Initial Video" and "Final Video": Always denote consecutive episodes (initial  
→ video i and final video i+1).
5. Provide exactly five multiple-choice options (A-E). Option E should be reserved  
→ for responses like "None of the above" or "All of the above," and should be  
→ correctly used occasionally to ensure difficulty.
6. False Answers (A-D):
  - Use only objects explicitly mentioned within the provided CSV file.
  - For other CSV types ("Object Location Change", "Object Detection", "Object  
→ State Change"), only use unique objects stated within their respective CSV  
→ files.
7. Do not explicitly state the video numbers within the question text itself.
8. If multiple changes occur within the same episode interval (e.g. 2 or 3 objects  
→ changed location), create multiple distinct questions ensuring each has only  
→ one correct answer.
9. Language and tone of the questions should feel conversational, natural, and  
→ realistic-similar to how a human would genuinely query a VLM about scene  
→ changes. Avoid robotic or unnatural phrasing.

10. Difficulty Levels:
- Level 1 (hardest): Do not mention the specific object or the change  
 ↳ explicitly. Example:  
 > *"Did anything change in the cubicle between these episodes?"*
  - Level 2 (medium): Specify the object explicitly, ask about the type of change. Example:  
 ↳ change. Example:  
 > *"There was a change in the number of steel mugs in the cubicle. What exactly changed?"*  
 > *"The mouse in the cubicle has been touched, what happened to it?"*
  - Level 3 (easiest): Clearly state the type of change and ask about the specific object. Example:  
 ↳ specific object. Example:  
 > *"An object's quantity increased between the episodes. Which object increased in number?"*
- Ensure a balanced mix of difficulty levels unless otherwise instructed.

## ## Examples

1. Object Detection (Medium):
- > Question: *"The Roman head statue disappeared from the cubicle. Where was it located before disappearing?"*
  - > Multiple Choice:
  - > - A: *On top of the PC*
  - > - B: *In front of the left monitor*
  - > - C: *Beside the red book*
  - > - D: *On the desk*
  - > - E: *None of the above*
2. Object Detection (Easy):
- > Question: *"A new object appeared on the desk. Which object appeared?"*
  - > Multiple Choice:
  - > - A: *Eiffel Tower toy*
  - > - B: *Green triceratops toy*
  - > - C: *Blue exploding kittens board game*
  - > - D: *Butter cookies metal box*
  - > - E: *None of the above*
3. Object Detection (Hard):
- > Question: *"Did anything appear or disappear between these episodes? If yes, what specifically changed?"*
  - > Multiple Choice:
  - > - A: *Eiffel Tower toy appeared; stapler disappeared*
  - > - B: *Roman head statue disappeared*
  - > - C: *Green triceratops toy appeared*
  - > - D: *Butter cookies metal box disappeared; green triceratops toy appeared*
  - > - E: *None of the above*

Listing 4: Example of the prompt feed into Gemini for the creation of the object detection questions of the local video changes.

## 2.4 Prompt for the Spatial Association VQA Benchmark

To evaluate the Spatial Association VQA Benchmark, Gemini 2.5 Pro [1] was given one global video in combination with the prompt shown in Listing 5.

Provide me with a count of the number of cubicles seen in the video and list the  
 ↳ cubicles with their id and name.

Listing 5: Prompt utilized in Spatial Association VQA Benchmark

## 2.5 Prompt for Single/Multi-Cubicle-Multi-Temporal VQA Benchmark

To benchmark Gemini 2.5 Pro [2] on the Single and Multi-Cubicle-Multi-Temporal VQA Benchmark, each question was appended to the prompt shown in Listing 6 and passed to the model.

You are taking a multiple-choice benchmark.

To answer the questions, you will be provided with two videos. The first video is  
→ the initial state of the scene, and the second video is the final state of the  
→ scene. Please pay attention to the changes of objects in the scene between the  
→ two videos.

Furthermore, if we say an object had been removed, vanished, or disappeared from a  
→ cubicle it means it will not be present in the second video. If we say an  
→ object has appeared in a cubicle it means it will be present in the second  
→ video. If we say an object has been moved, it means it has moved within the  
→ cubicle or between cubicles.

Please answer the question to best of your ability.

For each question below, reply with the single letter (A-E) that you believe is  
→ correct. Do not provide explanations-only the letter.

Please return a json object in the following format:

```
"answers": [
{
  "question": "Q1",
  "answer": "B"
},
{
  "question": "Q2",
  "answer": "C"
},
{
  "question": "Q3",
  "answer": "C"
},
{
  "question": "Q4",
  "answer": "B"
},
{
  "question": "Q5",
  "answer": "C"
},
{
  "question": "Q6",
  "answer": "C"
},
{
  "question": "Q7",
  "answer": "C"
}
....
]
```

Please use the Q# corresponding to the questions provided.  
Also please always answer the questions, never return an empty json.

Listing 6: Prompt used for Global - Video Question Answering Benchmark.

### 3 Temporal Change Alignment Example

[Listing 7](#) illustrates temporal change alignment for cubicle Alana 2038Q between Episode 5 and Episode 6. Each human-annotated event is semantically matched to a corresponding VLM-generated event based on object descriptions. The resulting alignments are categorized into three groups:



*Matched Change (True Positive), Only in Output (False Positive), and Only in Ground Truth (False Negative).*

```
{
  "Matched Change": {
    "C1": {
      "Output": {
        "Object": "small red and white items",
        "Change Type": "Object Counting",
        "Change Detail": "count changed from <1> to <2>"
      },
      "Ground Truth": {
        "Object": "tea bags",
        "Change Type": "Object Counting",
        "Change Detail": "count changed from 1 to 3"
      }
    },
    "C2": {
      "Output": {
        "Object": "light green hoodie",
        "Change Type": "Object Location Change",
        "Change Detail": "moved from <draped over the back of the office
↪ chair> to <on the right-hand section of the desk surface>"
      },
      "Ground Truth": {
        "Object": "green jacket",
        "Change Type": "Object Location Change",
        "Change Detail": "moved from on top of chair to right side of
↪ table"
      }
    }
  },
  "Only in Output": {
    "C1": {
      "Object": "yellow pencils",
      "Change Type": "Object Location Change",
      "Change Detail": "moved from <lying horizontally on the desk surface,
↪ to the right of the main keyboard and next to the clear plastic
↪ holder> to <placed vertically inside the clear plastic holder>"
    }
  },
  "Only in Ground Truth": {
    "C1": {
      "Object": "pair of chopsticks",
      "Change Type": "Object Detection",
      "Change Detail": "appeared at left side of table"
    },
    "C2": {
      "Object": "picture frame",
      "Change Type": "Object State Change",
      "Change Detail": "state changed from photo holder on the frame to
↪ photo holder removed"
    }
  }
}
```

Listing 7: Example of the temporal change alignment on cubicle *Alana 2038Q* between episode 5 and episode 6.

## 40 4 Sample Questions

41 [Table 3](#) and [Table 4](#) present example question sets generated by Gemini (2.5 Flash [1] and 2.5 Pro  
42 [2]) for the Multi-Cubicle-Multi-Temporal VQA (global change questions) and the Static Associa-

tion–Semantic Mapping VQA, respectively. The Single-Cubicle-Multi-Temporal VQA (local change questions) follows the same generation pattern as the multi-cubicle variant.

Table 3: Global multiple-choice questions automatically generated by Gemini.

| Category         | Example Question (Options A–E)  |
|------------------|---|
| Object Counting  | <i>“In Jack. J’s cubicle, what object’s quantity has been changed?”</i><br>A) Steel mug from 2 to 1; B) Plastic water bottle from 1 to 3; C) Led pencils from 1 to 5; D) Increased from 0 to 1; E) None of the above  |
| Object Detection | <i>“Which object appeared on Alexandria’s cubicle?”</i><br>A) Tim Hortons donut box; B) Coffee kettle; C) Roman-head toy statue; D) Colourful backpack; E) None of the above  |
| Object Location  | <i>“Where has the Triceratops been moved to, it was last seen in Emily’s cubicle, where is it now?”</i><br>A) Amy’s cubicle, next to the acne patch bag; B) Jack S.’s cubicle, in front of the monitors.; C) Jack S.’s cubicle, next to the mug beside the monitor.; D) Alana’s cubicle, on top of the PC tower. E) None of the above |
| Object State     | <i>“In Frank’s cubicle, what happened to the Laptop?”</i><br>A) The Laptop remained closed; B) The Laptop was moved to the shelf; C) The White flower bag was placed on the Laptop; D) The Laptop’s screen brightness was increased; E) None of the above   |

Table 4: Static Association–Semantic Mapping question testing cubicle/room location.

| Category              | Image ID | Example Question (Options A–E)  |
|-----------------------|----------|---|
| Cubicle/Room Location | 001138   | <i>“How would you describe the location of Frank’s cubicle relative to the visible doors?”</i><br>A) Between Door 2033 and Door 2035; B) Opposite Door 2033; C) Adjacent to the wall with Door 2033 and Door 2035; D) Far down the hallway from the doors; E) None of the above |

## 5 Keyframe Extraction

To extract the keyframes for [Section 2.1](#), all videos are first downsampled to 5 frames per second and longer videos are split into 5-minute segments. Each segment includes a 30-second overlap with the previous one to provide a warm-up period for stable initialization.

Keyframes were then selected using MAST3R-SLAM [3], a recent state-of-the-art monocular dense SLAM system designed for real-time operation. MAST3R-SLAM leverages two-view 3D reconstruction priors from MAST3R to jointly solve for tracking, mapping, and loop closure, even under challenging conditions such as time-varying or unknown camera models. A new keyframe is inserted when tracking quality degrades, specifically when the number of valid feature matches or the number of unique matched keyframe pixels falls below a threshold. This ensured that our keyframes were both geometrically informative and temporally consistent for downstream processing.

## 6 Expanded Results

This section presents the complete experimental results from the main paper and supplements them with additional qualitative examples that highlight Gemini’s points of weakness.

### 6.1 Spatial Association VQA

As discussed in the main paper, we evaluated Gemini 2.5 Pro’s ability to resolve spatial associations in dynamic office scenes using the six global-change videos. For each episode, the model was prompted to count the number of visible cubicles and to list every cubicle’s numeric ID together with its occupant’s name. Counting accuracy was measured with the mean absolute percentage error

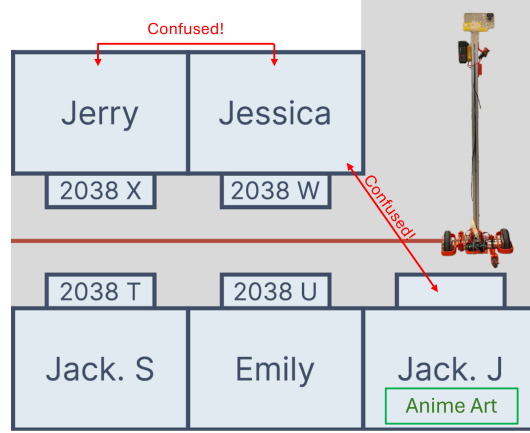


Figure 1: Enter Caption

(MAPE), while listing performance was assessed with precision and recall, averaged across episodes. The full results, summarised in Table 5, indicate consistently weak performance: the model never produces the correct cubicle count (average MAPE 27.5 %) and achieves mean precision and recall below 0.50.

A typical misstep in Episode 0 illustrates how Gemini’s spatial reasoning fails so frequently. As shown in Fig. 1 Jessica cubicle’s ID is 2038 W. However, what Gemini returns when given episode 0 and asked to list every cubicle’s numeric ID together with its occupant’s name is "Jessica’s cubicle ID with the anime drawing on the whiteboard is 2038 X". Not only is the ID wrong with that ID belonging to Jerry, but anime drawing on the whiteboard is not on Jessica’s cubicle but in Jack. J’s cubicle instead. The error reflects a broader pattern: Gemini latches onto a salient object but then assigns that object and the associated cubicle ID, and owner to the wrong location, leading to inflated counts and incorrect ID-to-occupant matches across episodes.

Table 5: Gemini 2.5 Pro [2] (Temperature ( $T=0.0$ )) answers for cubicle counting and listing tasks for global change videos in which the ground truth number of cubicles is 23.

| Episode        | Counting | Counting MAPE | Listing Precision | Listing Recall |
|----------------|----------|---------------|-------------------|----------------|
| Episode 0      | 28       | 21.7%         | 0.464             | 0.565          |
| Episode 1      | 12       | 47.8%         | 0.667             | 0.348          |
| Episode 2      | 16       | 30.4%         | 0.563             | 0.391          |
| Episode 3      | 16       | 30.4%         | 0.563             | 0.391          |
| Episode 4      | 18       | 21.7%         | 0.500             | 0.391          |
| Episode 5      | 26       | 13.0%         | 0.192             | 0.278          |
| <b>Average</b> | -        | 27.5%         | 0.491             | 0.394          |

## 6.2 Temporal Association VQA

Detecting changes in real office environments is essential for tasks such as anomaly detection (e.g., missing items) and object tracking (e.g., locating a misplaced phone). To mirror these use cases, we assessed Gemini 2.5 Pro’s ability to recognise and reason about changes over time. The model was given pairs of local videos showing the same cubicle at different timestamps and asked to identify any differences, returning its output in JSON format. Human annotations capturing the true changes served as the ground truth.

We aligned each human-annotated event with the corresponding model output by semantically matching object descriptions. Events were then categorised as *Matched Change* (true positive), *Only in Output* (false positive), or *Only in Ground Truth* (false negative). Listing 7 illustrates this alignment.

87 In that example, Gemini correctly flagged the addition of tea bags and the relocation of a green hoodie  
88 but missed the introduction of a pair of chopsticks and the removal of a photograph in a picture frame.  
89 It also hallucinated the movement of pencils into a clear plastic holder. Thus, it missed two of four  
90 real changes and reported one spurious change. The quantitative results in Table 6 echo this pattern:  
91 an overall F1 score of 0.52 indicates limited effectiveness in temporal change detection.

Table 6: Performance on the *Temporal Association VQA* task.

| Statistic | Matched (TP) | Only in GT (FN) | Only in Output (FP) | Precision | Recall | F <sub>1</sub> |
|-----------|--------------|-----------------|---------------------|-----------|--------|----------------|
| Values    | 587          | 412             | 667                 | 0.47      | 0.59   | 0.52           |

### 92 6.3 Single-Cubicle-Multi-Temporal VQA

93 For each temporal episode we supply Gemini with the two walk-through videos of a single cubicle,  
94  $\langle v_{e-1}, v_e \rangle$ , together with the multiple-choice questions generated from that cubicle’s local-change  
95 log. Queries follow the structured prompt shown in the supplementary material; JSON output is  
96 enforced through Gemini’s structured-response schema.

97 **Prompting.** The supplementary-material template is used with temperature  $T=0.0$  to obtain deter-  
98 ministic output; if Gemini returns invalid JSON, we retry once at  $T=0.25$ .

99 **Video pre-processing.** Videos remain at their native 1080 p resolution, but audio tracks are stripped  
100 and the frame rate is down-sampled to 10 fps to reduce file size.

101 **Scoring.** The model’s JSON answers are compared with ground-truth keys, and accuracy is computed  
102 for each change category as well as overall.

103 Table 7 summarises the per-cubicle results. Gemini attains an overall accuracy of 56.8 %, with  
104 marked variability across cubicles, suggesting that certain workspaces are more challenging than  
105 others. By category, Object Detection is most reliable (63.6 %), followed by Location (61.9 %), State  
106 (53.1 %), and Counting (48.6 %). These trends are consistent with expectations: precise counting  
107 and subtle state discrimination (e.g., lid-open vs. lid-closed) require finer spatial-temporal resolution  
108 than simply recognising or localising objects.

Table 7: Accuracy (%) of Gemini 2.5 Pro [2] on the *Single-Cubicle-Multi-Temporal VQA* task. Asterisks (\*) denote runs that required a higher sampling temperature ( $T=0.25$ ) to obtain valid JSON output; all other runs used  $T=0.0$

| Cubicle                   | Total | Object<br>Dete-<br>ction | Object<br>Loca-<br>tion | Object<br>Count-<br>ing | Object<br>State |
|---------------------------|-------|--------------------------|-------------------------|-------------------------|-----------------|
| Alana-2038Q               | 45.0% | 60.0%                    | 56.0%                   | 20.0%                   | 44.0%           |
| Alexandria-2008M*         | 43.6% | 53.8%                    | 44.0%                   | 40.0%                   | 28.0%           |
| Amin-2008E*               | 55.6% | 61.5%                    | 52.0%                   | 56.0%                   | 52.2%           |
| Daniel-2038Y*             | 57.3% | 70.8%                    | 70.8%                   | 38.5%                   | 50.0%           |
| Emily-2038U*              | 57.6% | 72.0%                    | 76.0%                   | 52.0%                   | 29.2%           |
| Fernando-2041S*           | 65.6% | 72.0%                    | 80.0%                   | 50.0%                   | 60.0%           |
| Jack S-2038T*             | 45.5% | 48.1%                    | 41.7%                   | 24.0%                   | 68.0%           |
| Jason-2008K*              | 64.7% | 73.1%                    | 53.8%                   | 65.4%                   | 66.7%           |
| Josh-2008S*               | 72.6% | 66.7%                    | 62.5%                   | 73.9%                   | 87.5%           |
| Kristine-2038N*           | 63.6% | 60.0%                    | 83.3%                   | 68.0%                   | 44.0%           |
| <b>Mean</b>               | 56.8% | 63.6%                    | 61.9%                   | 48.6%                   | 53.1%           |
| <b>Standard Deviation</b> | 9.4%  | 8.1%                     | 14.1%                   | 17.2%                   | 17.4%           |

## 109 **References**

- 110 [1] Google DeepMind. Gemini 2.5 Flash Preview: Price–Performance Optimised Multimodal Model. [https://ai.google.dev/gemini-api/docs/models#2\\_5\\_flash\\_experiment](https://ai.google.dev/gemini-api/docs/models#2_5_flash_experiment), 2025. Accessed: 22 May  
111 2025.  
112
- 113 [2] Google DeepMind. Gemini 2.5 Pro Preview: Multimodal Large Language Model. [https://ai.google.dev/gemini-api/docs/models#2\\_5\\_pro\\_experiment](https://ai.google.dev/gemini-api/docs/models#2_5_pro_experiment), 2025. Accessed: 22 May 2025.  
114
- 115 [3] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MAST3R-SLAM: Real-time dense SLAM with 3D  
116 reconstruction priors. *arXiv preprint*, 2024.