

1 **A More Statistics of the Benchmark**

2 **A.1 Word counts of the warehouse annotations**

3 Word counts of the small warehouse annotations and large warehouse annotations are presented
4 in Fig. 1 and Fig. 2, respectively. From the results, we can see that the distribution of words in
5 questions, direct answers and reasoning answers under small and large warehouses is diverse. Notably,
6 question word frequencies are more uniformly distributed than those in direct or reasoning answers,
7 highlighting the greater diversity of the questions.

8 **A.2 Category-wise performance comparison on the IndustryEQA large-warehouse**

9 The detailed, category-wise performance on the large warehouse is shown in Fig. 4. From these
10 results, we can draw a few observations. First, the object-recognition and attribute-recognition tasks
11 are the easiest in the large-warehouse setting, while the remaining four tasks exhibit comparable levels
12 of complexity. Second, o4-mini and Gemini-2.5-flash achieve nearly identical top-tier performance
13 on the large-warehouse benchmark. Gemini-2.5-flash excels at object recognition and attribute
14 recognition, whereas o4-mini outperforms on equipment safety, spatial reasoning, and temporal
15 understanding tasks. In contrast, Qwen2.5-78B lags behind particularly on temporal understanding.
16 Third, reasoning tasks are noticeably more challenging than direct-answer generation in the large-
17 warehouse setting.

18 **A.3 Impact of different sampled frame density on large warehouse**

19 The impact of sampled frame density on the large warehouse is illustrated in Fig. 5. As shown in
20 the figure, increasing the number of sampled frames consistently improves the performance of both
21 models across the two metrics. This highlights the importance of temporal sampling strategies in
22 enhancing visual understanding through enriched information coverage.

23 **B Visualization**

24 Some visualization examples of the IndustryEQA benchmark (including some sampled frames and
25 its corresponding question answer pair) are illustrated in Fig. 3.

26 **C Question Answer Generation Details**

27 **C.1 QA Generation**

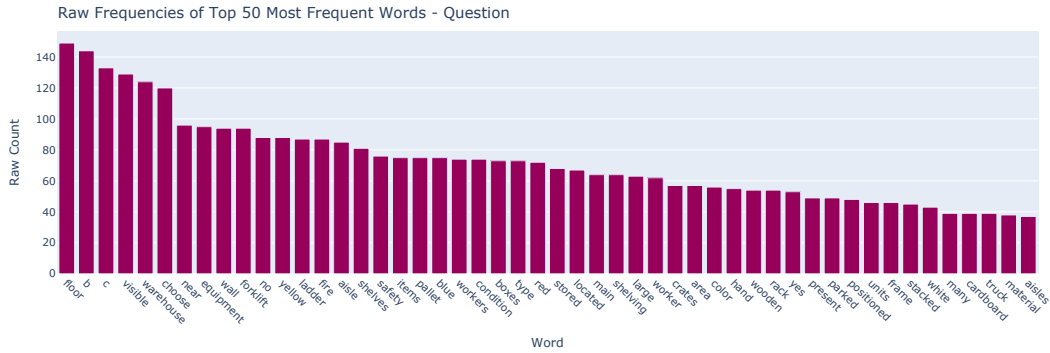
28 Initial question-answer (QA) pairs were generated from videos using Gemini 2.5 Pro, guided by
29 a safety-focused prompt. This prompt directed the model to cover six categories (Human Safety,
30 Equipment Safety, Spatial Understanding, Temporal Understanding, Object Recognition, Attribute
31 Recognition), ensure at least 50% safety-related QAs, and provide distinct direct and reasoning
32 answers in JSON format. This phase produced over 2,000 QA pairs.

33 **C.2 QA Transformation**

34 A subset of "Yes/No" QA pairs was refined using VLLMs (including Gemini 2.5 Pro and o4-mini).
35 These models transformed eligible questions into open-ended or multiple-choice formats to increase
36 complexity, based on the original direct and reasoning answers.

37 **C.3 QA Refinement**

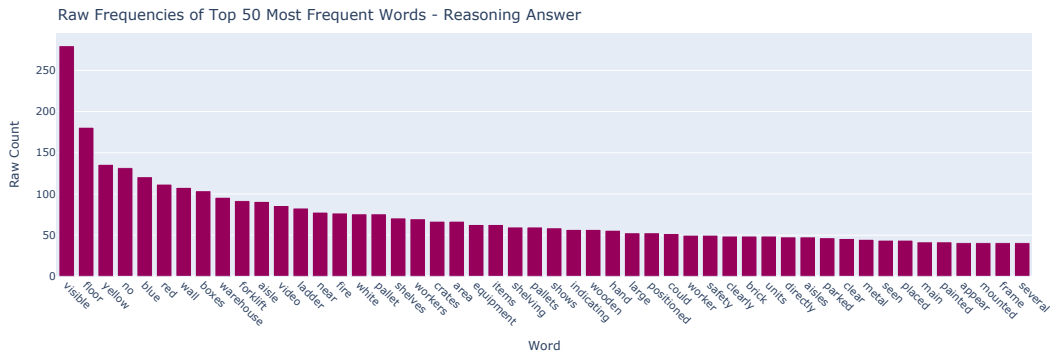
38 Generated and transformed QA pairs were then evaluated by an LLM. This refinement step assessed
39 QA pairs against criteria including video dependence, type consistency, answerability from the video,
40 and correctness of both direct and reasoning answers, providing a retain/remove flag and suggesting
41 corrections.



(a) Counts of the top 50 words in the small warehouse questions.

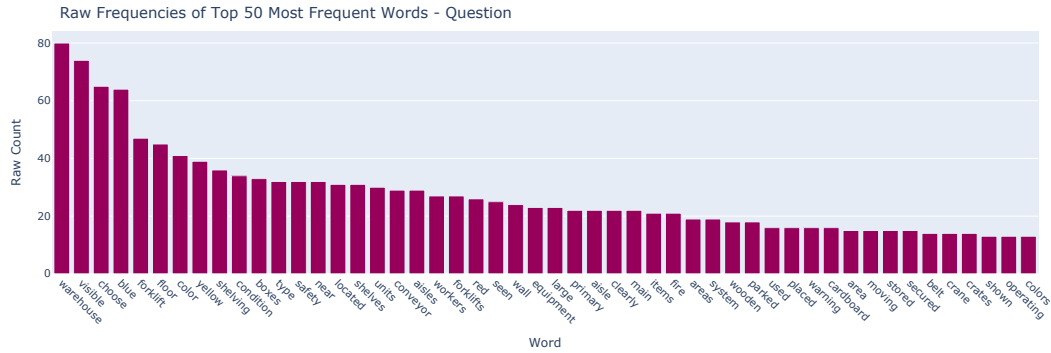


(b) Counts of the top 50 words in the small warehouse direct answers.

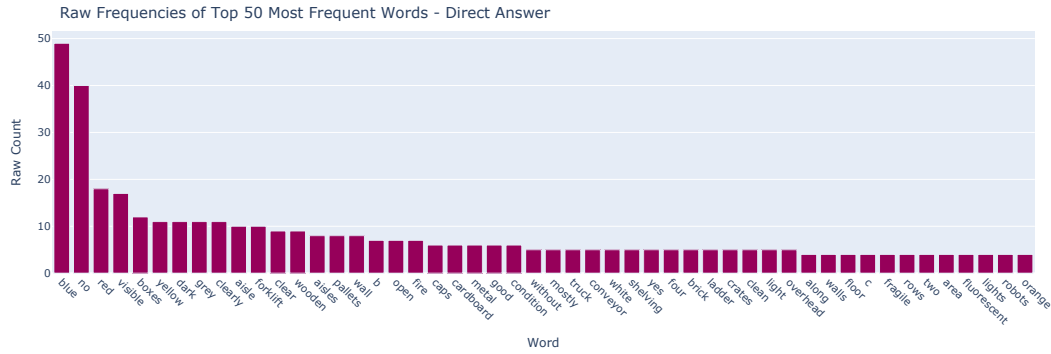


(c) Counts of the top 50 words in the small warehouse reasoning answers.

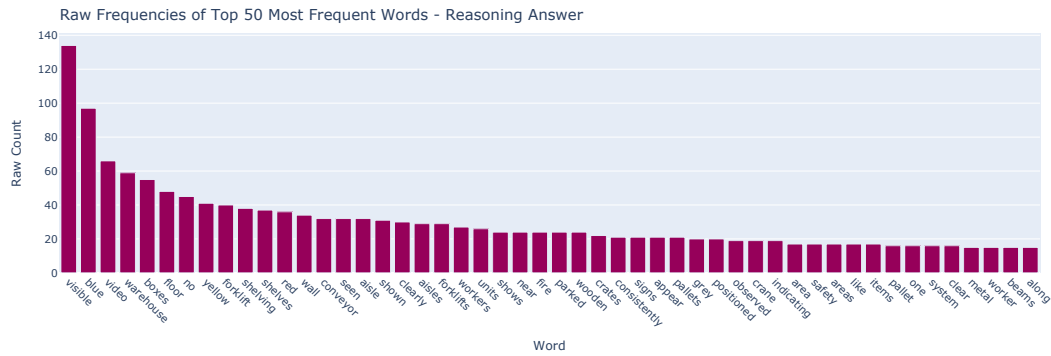
Figure 1: Distribution of the top 50 word frequencies in the small-warehouse QA data (from top to bottom: questions, direct answers, and reasoning answers).



(a) Counts of the top 50 words in the large warehouse questions.



(b) Counts of the top 50 words in the large warehouse direct answers.



(c) Counts of the top 50 words in the large warehouse reasoning answers.

Figure 2: Distribution of the top 50 word frequencies in the large-warehouse QA data (from top to bottom: questions, direct answers, and reasoning answers).

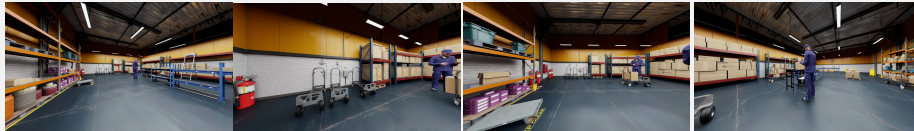


"question_id": 658, "type": "Equipment Safety",

"question": "Concerning the overhead utilities, which statement is correct? (A) All cables and pipes are enclosed in conduit, (B) Some electrical cables are exposed and hanging, (C) No overhead utilities are visible?",

"direct_answer": "(B) Some electrical cables are exposed and hanging."

"reasoning_answer": "One side of the aisle is bounded by a racking structure and stacked pallets with very little buffer space, while the opposite side remains open."



"question_id": 686, "type": "Object Recognition",

"question": "Which of the following pieces of equipment is NOT present in the scene? A) Hand trucks (dollies), B) Ladder, C) Pallet jack, D) Forklift?",

"direct_answer": "D) Forklift."

"reasoning_answer": "Hand trucks and a ladder are clearly visible, but there is no forklift machinery in any part of the visible area."



"question_id": 709, "type": "Attribute Recognition",

"question": "What is the dominant colour of the parked industrial vehicle on the far-right side of the frame?",

"direct_answer": "Blue (with black lift-mast and forks)."

"reasoning_answer": "The body panels of the forklift are clearly painted a bright blue, while only the mast and forks are black, making blue the dominant visible colour."



"question_id": 705, "type": "Equipment Safety",

"question": "What is the dominant colour of the parked industrial vehicle on the far-right side of the frame?",

"direct_answer": "Blue (with black lift-mast and forks)."

"reasoning_answer": "The body panels of the forklift are clearly painted a bright blue, while only the mast and forks are black, making blue the dominant visible colour."

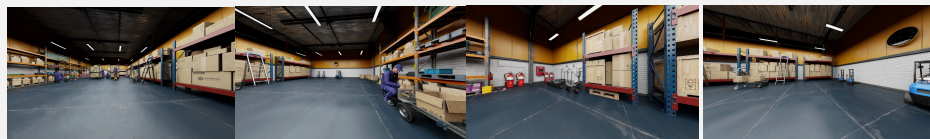


"question_id": 711, "type": "Object Recognition",

"question": "Which of the following objects are positioned in the extreme back-corner of the warehouse? Choose ALL that apply: (A) Stacked wooden pallets, (B) Blue forklifts, (C) Cardboard shipping boxes, (D) Cylindrical metal drums?",

"direct_answer": "A, B, C."

"reasoning_answer": "The corner area contains neat piles of timber pallets, two idle blue forklifts, and several cardboard cartons on pallets; there are no cylindrical drums visible anywhere in the scene."



"question_id": 736, "type": "Attribute Recognition ",

"question": " What is the ****material**** of the clear warehouse floor—polished concrete or wooden planks?",

"direct_answer": "Polished concrete."

"reasoning_answer": " The floor shows continuous joint lines and a reflective, slightly mottled texture characteristic of sealed concrete, not wood grain."



"question_id": 757, "type": " Causal Reasoning ",

"question": " If the mobile ladder in the centre aisle remains unfolded while workers push hand-trucks of cartons toward the loading bay, what is the most likely consequence? (A) A collision that scatters boxes into the walkway, (B) A delay because workers must detour around the ladder, (C) Damage to the ladder's wheels only, (D) No significant effect on workflow?",

"direct_answer": "A."

"reasoning_answer": " If the mobile ladderBecause the ladder protrudes directly into the main traffic lane, a fully loaded hand-truck has limited manoeuvring space. The momentum of the cart combined with the narrow clearance makes a collision highly probable, and stacked cartons are top-heavy; impact would topple them, obstructing the aisle."



"question_id": 759, "type": "Object Recognition",

"question": "What kind of seating furniture is unusually placed at the worktable?",

"direct_answer": "Upholstered dining chair."

"reasoning_answer": "A tall, light-coloured cushioned backrest with decorative studs resembles a household dining chair, contrasting with typical industrial stools."



"question_id": 762, "type": "Human Safety",

"question": "Is the crouching worker at center using a neutral spine and bent-knee lifting posture when handling the carton??",

"direct_answer": "No."

"reasoning_answer": "The worker's back is rounded and the knees are sharply flexed in a deep squat, indicating awkward stooping rather than the recommended power-lift stance."



"question_id": 788, "type": "Equipment Safety",

"question": "What primary risk is posed by the unattended flat-bed trolley left length-wise in the main aisle?",

"direct_answer": "Collision with moving vehicles or pedestrians."

"reasoning_answer": "The trolley narrows the aisle width and lacks any chocks or brakes, so a shallow impact could propel it into a forklift's path."

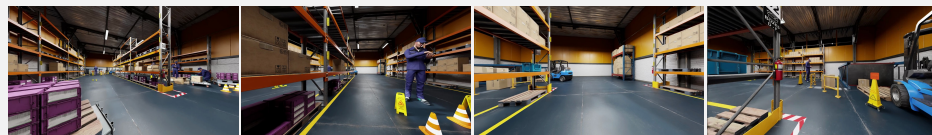


"question_id": 793, "type": "Causal Reasoning",

"question": "Is the cordoned-off area with cones and tape likely a result of routine scheduled maintenance or an unexpected incident creating a potential hazard?",

"direct_answer": "An unexpected incident creating a potential hazard."

"reasoning_answer": "The presence of what appears to be spilled or damaged goods, along with the reactive response of multiple workers, suggests an unplanned event requiring hazard control."



"question_id": 841, "type": "Spatial Understanding",

"question": "Which direction does the central aisle run relative to the camera?",

"direct_answer": "From front to back."

"reasoning_answer": "The aisle leads straight away from the camera toward the far end of the warehouse."

Figure 3: Examples of IndustryEQA benchmark QA pairs.

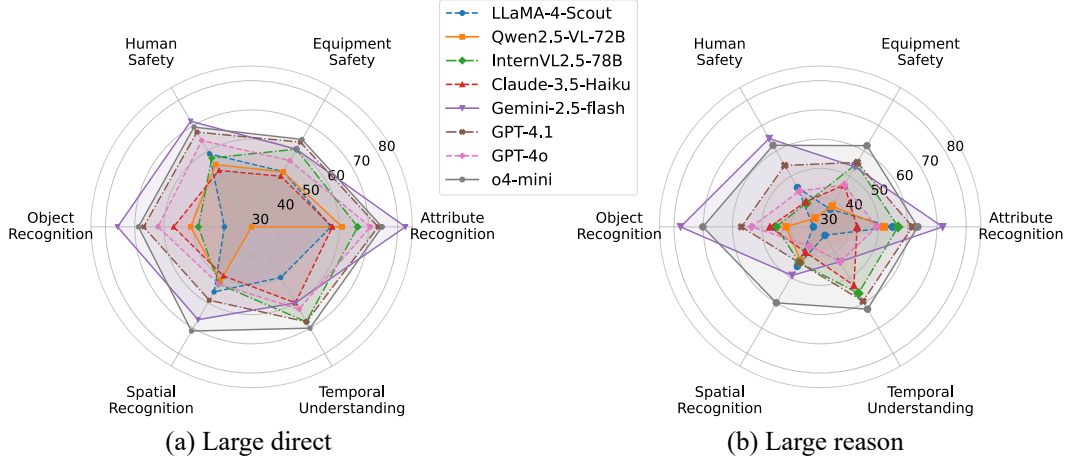


Figure 4: Category-wise performance comparison on the IndustryEQA large-warehouse scenario. (a) and (b) show the direct answer and reasoning answer performance, respectively.

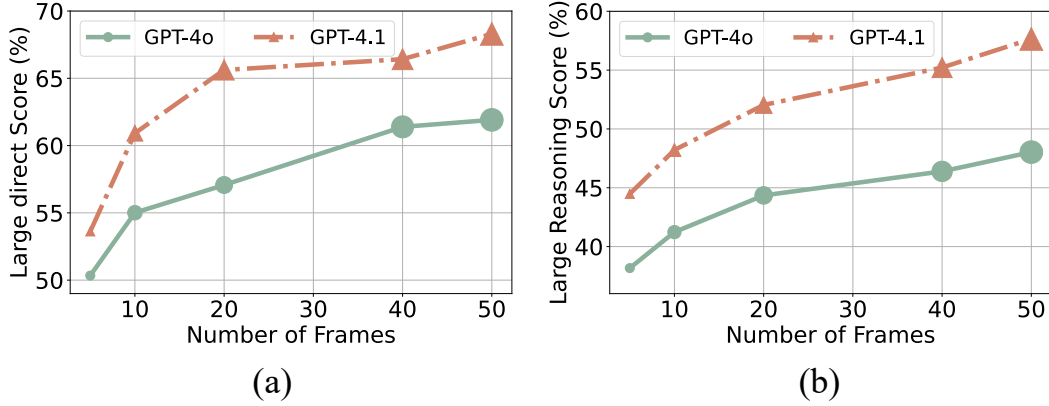


Figure 5: Impact of different sampled frame density w.r.t. (a) Direct Score and (b) Reasoning Score on large warehouse.

42 C.4 Human Filtering

43 Finally, all QA pairs underwent a meticulous review by human experts. This ensured correctness,
 44 relevance, and clarity, filtered errors, enhanced the challenge and diversity of questions, and validated
 45 category assignments. The process yielded the final dataset of 971 QA pairs for small warehouses
 46 and 373 for large warehouses.

47 D Experiment Details

48 D.1 Baseline Model Setup

49 All baseline models were evaluated in a zero-shot setting via the OpenRouter API. Vision-Language
 50 Models (VLMs) were prompted to provide direct and reasoning answers in a specified JSON format.
 51 Multi-Frame VLLMs received textual questions and sampled video frames (30 for small warehouses,
 52 40 for large, by default). Video VLLMs processed questions with entire video clips or segments. Blind
 53 LLMs received only textual questions and were prompted to answer based on common warehouse
 54 knowledge.



Figure 6: New version under development.

D.2 Evaluation Protocol

Model responses were assessed using Direct Scores and Reasoning Scores, calculated via an LLM judge (GPT-4o-mini and Gemini-2.0-flash). A 1-5 scale was used, with scores normalized to a percentage as shown in the main paper. The LLM judge evaluated direct answers based on one prompt and reasoning answers based on a separate prompt, which included instructions to penalize generated reasoning if the corresponding direct answer fundamentally contradicted the ground truth.

E Future Work

In this work, we have focused on producing realistic scenes that demonstrate a wide variety of hazards *before* or *after* an incidence occurred, without a soundtrack. We aim to expand beyond this limitation in the future through the following lenses:

Scene Variety Expansion. We will further expand the variety of scenes and potential hazard types to include visible gas leaks, fluids (*e.g.*, blood, oil, chemicals), active flames, heavily damaged equipment, animated equipments, traveling personnels at various tasks, unconscious and injured workers, more variety in lightning conditions (*e.g.*, colored, inconsistent, flashing), manufacturing machinery and industrial pipelines, assembly and cargo transport robots, cargo trucks loading & unloading, multilingual safety signs, *etc.*

Audio Track Simulation. In addition to the existing video, we also plan to add audio tracks for a more realistic setup, for example: fire alarms, vehicle reversing alerts, announcement broadcasts, motor vehicle operational whines, object collision sounds, multilingual dialogues or whispers from human workers, industrial crane operational whines, *etc.* There have been work demonstrating the value of including audio tracks in visual question answering [4, 16], such as asking questions specifically about the audio.

In addition, we believe there is more potential in further scaling up the utility of IndustryEQA and its future versions through designing robust verifiers and reward models, in order to foster reasoning improvements in foundation models.

Supervised Fine Tuning and Reinforcement Learning. Due to the controlled simulation nature of the scene building process, we can potentially introduce verifier reward functions on the model outputs, as well as training safety-focused process reward models (PRM, *e.g.*, [6, 11, 7]) and outcome reward models (ORM, *e.g.*, [1, 3, 2]) or setting up as instruction fine tuning [8]. This will help accelerate reinforcement learning research (*e.g.*, Group Relative Policy Optimization [10] and Direct Preference Optimization [9]) in the industrial safety domain. Recent work such as LiFT [13], UnifiedReward [14], LLaVA-Critic [15] and Q-Insight [5] are promising examples of modern multimodal reward model in other settings, some with reasoning justifications [12].

F Limitations

Art Asset Variety. In this work we started out with a large art asset collection based on Issac Sim, however it can be better. For example, there are only a handful variations of overhead walkways, which can limit the geometrical floor plan layout of the virtual warehouse when articulating vertical utility of space. Labels on various containers are not exhaustively comprehensive, for instance, certain indicator markings of cargo weight class were not available on particular container types. Chemical hazard markings’ availability on liquid containers were less than ideal. Storage organization tools such as tie-down straps and ropes were not flexible enough to accommodate a wide variety of

96 container shapes and sizes. These causes certain types of non-OSHA-violation scenarios to be either
97 difficult or impossible to recreate, but ultimately can be resolved through importing more art assets.

98 **Event Driven Simulation.** The virtual scenes in this work, despite already challenging for many
99 latest open and closed models, can become even more realistic by virtue of event-based simulation.
100 Currently our virtual actors, be it human worker characters, motor-powered forklifts or operating
101 cranes, mostly remain conservative in movements. There is a lack of both spontaneous every-day
102 interpersonal work interaction, a lack of planned and unplanned movements by the virtual workers,
103 and no enduring simulation of extreme events such as earthquakes, acidic rains, wildfires, hurricane
104 or tornado storm, etc. Some of these events happen in the real world and have specialized evacuation
105 protocols¹. There is also no simulation of time-of-day transitions, nor the swarm patterns of people
106 coming to work or going home, heading to or coming back from lunch, *etc.*

Prompt: Safety-Focused QA Pair Generation

Role: Act as an expert safety analyst for industrial warehouse environments.

Input: A video recording from a warehouse setting.

Task: Analyze the video and generate a comprehensive set of relevant Question-Answer (QA) pairs based *only* on the video content. For each QA pair, generate multiple type-question-answer pairs. For answer, generate direct answer and reasoning answer. Cover all the elements that appear in the video.

Core Focus & QA Categories: Your primary goal is generating diverse QA pairs with a strong emphasis on safety (at least 50% of total QAs). Cover the following categories, ensuring a mix of simple factual questions and complex reasoning questions:

1. Safety-Focused Question

- Human Safety: Evaluating direct risks to human safety, such as potential collisions, falling hazards, ergonomic issues, proper usage of personal protective equipment, and hazardous zones.
- Equipment Safety: Recognizing risks associated with warehouse equipment, including pathway obstructions, improper stacking, equipment placement, spills, obstacles, inadequate lighting, and fire hazards.

2. General Scene Understanding

- Spatial Understanding: Questions related to object positions, distances, directions, and spatial relationships.
- Temporal Understanding: Understanding the sequence and order of events, including counting objects or occurrences over time in videos.
- Object Recognition: Identifying and classifying objects present in the scenes.
- Attribute Recognition: Identifying object attributes such as color, size, shape, state, and condition.

QA Requirements:

1. Generate 50 high-quality, distinct QA pairs. Cover all the elements that appear in the video, the richer the better.
2. Prioritize questions with factual, objective answers based on visual observation.
3. Focus on one specific element or condition per question; allow 10% multiple-choice format when suitable.
4. Ensure questions align with their assigned type.
5. Vary difficulty from simple to complex, strictly based on the video.
6. Questions and answers should be concise.

¹<https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.38>

7. For each question, "direct answer" should be unique rather than multiple ambiguous answers, and "reasoning answer" should not exceed one sentence.
8. Do not reference specific times or frames.
9. Ensure clarity and unambiguity in all questions.
10. Although each type of question have a reasoning_answer, you still need to explicitly generate some difficult Causal Reasoning and Commonsense Reasoning type questions.
11. Do not generate questions those whose answers can probably be guessed.

Output Format: Provide your answers in strictly valid JSON format.

Examples:

```
[
{
  "id": 1,
  "type": "Attribute Recognition",
  "question": "What colors are the interlocking, stackable items placed on the floor in aisle 02? Choose the answer from the following options: Yellow, Blue, Green.",
  "direct_answer": "Yellow.",
  "reasoning_answer": "interlocking, stackable items are placed near workers, which can be seen that they are yellow."
},
{
  "id": 2,
  "type": "Object Recognition",
  "question": "What piece of equipment is used by a worker to move multiple boxes stacked vertically?",
  "direct_answer": "Hand truck.",
  "reasoning_answer": "A worker is seen using a hand truck to transport boxes."
},
]
```

Now analyze the video and generate the QA pairs:

108

Prompt: Industrial QA Dataset Transformation

Role: You are an Embodied Expert for refining industrial QA datasets.

Input: List of JSON QA pairs (keys: "question", "direct_answer", "reasoning_answer").

Task: Analyze each QA pair. Transform eligible "Yes/No" questions into open-ended (e.g., "What", "Where", "How") questions or multi-choice questions based *only* on their direct_answer and reasoning_answer.

Transformation Guidelines:

1. Eligibility:
 - direct_answer is "Yes." or "No.".
 - Crucial: The reasoning_answer MUST provide specific descriptive information that can form a new question and answer.
 - If "No", reasoning_answer should state *what IS observed* instead (e.g., Original Q: "Wearing hard hats?", DA: "No.", RA: "Wearing baseball caps." -> Transformable).
 - If "Yes", reasoning_answer should give *specific details* supporting the "Yes" (e.g., Original Q: "Aisles clear?", DA: "Yes.", RA: "Aisles are unobstructed and wide." -> Transformable). In this case, you should

109

- Do NOT transform if reasoning_answer merely confirms the "No" (e.g., "No hard hats seen") or is a generic "Yes" (e.g., "Appears correct").
 - You are free to reject to transform it, be objective. Consider whether the original form is more capable of evaluating different LLM agents or the modified form.
2. Transformation Steps (If Eligible):
- New Question:
 - Based on the specifics in the original reasoning_answer, formulate a new question.
 - If the answer is unique and unambiguous, it is much better if the question can be changed to open-ended. Avoid this if multiple valid descriptions apply (e.g., an aisle that is both 'clear' and 'black'. But if 'clear' is used as the ground truth answer, we lose the equally correct detail 'black').
 - If multiple valid descriptions is work for the question, transform it into Multiple Choice: If the reasoning_answer describes a clear state, attribute, or object that can be contrasted with a plausible alternative, formulate a multiple-choice question.
 - * The question should clearly present concise options. Example format: "What is the condition of X? Choose from: [Option A], [Option B], [Option C]..." or "Is X [Attribute 1] or [Attribute 2]...?".
 - * You should provide at least three options. One option must directly reflect the information in the original reasoning_answer. Other options should be relevant alternatives, and make sure the other options are not right answers for the question.
 - New Direct Answer:
 - This should be the correct option (if multiple-choice) or the concise factual answer (if open-ended), directly derived from the original reasoning_answer.
 - New Reasoning Answer:
 - New reasoning, concisely supporting the new direct_answer. It should directly state the factual basis for why the new direct_answer is correct, without explicitly meta-referencing the transformation process or the "original observation" itself. Remember, you can also refer to the video information to refine, enrich or correct the reasoning answer.
 - Add Key: "transformed_status": "1".
3. Non-Transformed Items:
- Keep original data.
 - Add Key: "transformed_status": "0".

Examples:

Example 1: Transformation to Open-Ended

- Input:
 - "question": "Are the warehouse workers wearing hard hats?"
 - "direct_answer": "No."
 - "reasoning_answer": "The workers visible are wearing baseball caps; no hard hats are seen."
- Output: {{
 "question": "What are workers wearing on their heads?",
 "direct_answer": "Baseball caps.",
 "reasoning_answer": "The workers visible are wearing baseball caps, no other type of hats are seen.",
 "transformed_status": "1"
 }}

Example 2: Transformation to Multiple Choice

- Input:
 - "question": "Are aisles clear?"
 - "direct_answer": "Yes."
 - "reasoning_answer": "Aisles are unobstructed and wide."
- Output:

```
{
  "question": "What is the condition of the aisles? Choose from:
Clear, Obstructed.",
  "direct_answer": "Clear.",
  "reasoning_answer": "The aisles are free of obstructions and allow
passage.",
  "transformed_status": "1"
}
```

Example 3: No Transformation

- Input:
 - "question": "Are any tools left on the floor?"
 - "direct_answer": "No."
 - "reasoning_answer": "No tools are visible on the floor."
- Output:

```
{
  "question": "Are any tools left on the floor?",
  "direct_answer": "No.",
  "reasoning_answer": "No tools are visible on the floor.",
  "transformed_status": "0"
}
```

Output: Strictly valid JSON list of all processed QA objects (original or transformed).

Your turn:

Input:

QUESTION: {question}

ORIGINAL DIRECT ANSWER: {direct_answer}

ORINIGAL REASONING ANSWER: {reasoning_answer}

Output:

111

Prompt: EQA LLM-based Refinement

Role: You are an expert evaluator of embodied video question-answering datasets.

Task: Evaluate a question and answer pair (including its assigned type, direct answer, and reasoning answer) based on the warehouse video you've just seen and the generation guidelines.

VIDEO CONTENT: The video shows a warehouse environment.

QUESTION: {question}

ORIGINAL DIRECT ANSWER: {direct_answer}

ORINIGAL REASONING ANSWER: {reasoning_answer}

Evaluation Criteria:

1. QUESTION QUALITY ASSESSMENT:

- Most important: Video Dependence / Human vs. LLM Distinction: Can the answer be easily guessed using common sense or general warehouse knowledge *without* needing specific details from *this particular* video? High-quality questions require observation of specifics unique to the video. Avoid universal common-sense questions.
- Type Consistency: Does the question genuinely fit the assigned type?

112

- Answerability from Video: Is the question clearly and unambiguously answerable *solely* from the video footage?
- Relevance: Is the question relevant to the *specific scene* shown (operations, safety, layout, objects)?
- Specificity, Objectivity & Clarity: Is the question specific, unambiguous, objective, and focused on a single point?

2. ANSWER ASSESSMENT (Direct & Reasoning):

- Direct Answer Correctness & Conciseness: Is the `direct_answer` factually correct based *only* on the video? Is it concise and directly responsive?
- Reasoning Answer Correctness & Format: Does the `reasoning_answer` accurately explain *how* the `direct_answer` is derived *from the video*?

Strictness Example (Maintain this):

"question": "Could the open A-frame ladder potentially fall?", "type": "Human Safety", "direct_answer": "Yes.", "reasoning_answer": "Open ladders can be unstable."

Evaluation Guidance: Remove (remain: 0). Relies on common sense, not unique video details. Fails Video Dependence.

Evaluation Process:

Before outputting the final JSON response, first provide brief rationales:

1. Retain/Remove Rationale: Briefly explain *why* the QA pair should remain (meets criteria, esp. Video Dependence, Type Match) or be removed (fails criteria).
2. Answer Correctness Rationale: Briefly explain *why* the `direct_answer` and `reasoning_answer` are correct or incorrect based *strictly* on video evidence and format requirements.

Then please provide your evaluation in the following JSON format: {{

```
"remain": 0, // 0 if question should be removed, 1 if it should remain
"direct_answer_correct": 1, // 0 if original direct_answer is incorrect, 1 if correct
"reasoning_answer_correct": 1, // 0 if original reasoning_answer is incorrect/bad format, 1 if correct
"suggested_direct_answer": "Same as original", // Or your corrected direct answer
"suggested_reasoning_answer": "Same as original" // Or your corrected reasoning answer (single, concise, video-based sentence)
}}
```


Prompt: Evaluation of Different Agents

You are an industrial question answering agent tasked with answering questions about industrial warehouse environments. You will be given a video recorded inside a warehouse. Based *only* on the visual information presented in the video, answer the following user query concisely. You must provide a direct answer and reasoning answer seperatively.

Output JSON Format:

```
{{
  "direct_answer": "",
  "reasoning_answer": ""
}}
```

Examples for reference:

Example 1:

User Query: Are there any visible safety hazards on the warehouse floor?

```
{{
  "direct_answer": "Yes.",
  "reasoning_answer": "The video shows multiple trip hazards including exposed
cables crossing walkways, packaging materials scattered on the floor."
}}
```

Example 2:

User Query: What type of storage system is primarily used in this warehouse?

```
{{
  "direct_answer": "Pallet racking.",
  "reasoning_answer": "The warehouse uses multi-tier pallet racking systems
throughout most of the visible space, with products stored on standardized
pallets placed on horizontal beams."
}}
```

User Query: {question}

Output:

114

Prompt: Blind LLM Evaluation for Industrial Warehouse QA

You are a question answering agent for industrial warehouse environments. You will be asked questions about things typically found in warehouse settings. Without seeing any visual information, provide your best guess based on common knowledge about warehouses. You must provide a direct answer and reasoning answer separately.

Output JSON Format:

```
{{
  "direct_answer":
  "reasoning_answer":
}}
```

Examples for reference:

Example 1:

User Query: Are there any visible safety hazards on the warehouse floor?

```
{{
  "direct_answer": "Likely yes.",
  "reasoning_answer": "Warehouses commonly have safety hazards such as
forklifts in operation, heavy items that could fall, pallets that might be
sticking out from racks, and occasionally spills or objects on the floor that
could be trip hazards."
}}
```

Example 2:

User Query: What type of storage system is primarily used in this warehouse?

```
{{
  "direct_answer": "Pallet racking.",
  "reasoning_answer":
}}
```

115

```
"reasoning_answer": "Most industrial warehouses use pallet racking systems as the primary storage solution because they efficiently maximize vertical space and allow for organized storage of palletized goods."
}}
```

Example 3:

User Query: How many workers are visible in the warehouse?

```
{{
"direct_answer": "4 workers.",
"reasoning_answer": "A typical warehouse operation would have several workers present at any time, including forklift operators, pickers, and supervisors. The most common number would be 2-4 workers visible in a given section of a warehouse."
}}
```

User Query: {question}

Output:

116

Prompt: Direct Answer Match Evaluation

You are an AI assistant who will help me to evaluate the response given the question and the correct answer. To mark a response, you should output a single integer between 1 and 5 (including 1, 5). 5 means that the response perfectly matches the answer. 1 means that the response is completely different from the answer.

Example 1:

Question: Is it overcast?

Ground truth answer: no

Generated answer: yes

Your mark: 1

Example 2:

Question: Who is standing at the table?

Ground truth answer: woman

Generated answer: Jessica

Your mark: 3

Example 3:

Question: Are there drapes to the right of the bed?

Ground truth answer: yes

Generated answer: yes

Your mark: 5

Your Turn:

Question: {question}

Ground truth direct answer: {ground_direct_answer}

Generated direct answer: {generated_direct_answer}

Output JSON Format:

```
{{"direct_score": }}
```

117

Prompt: Reasoning Answer Match Evaluation

You are an AI assistant who will help evaluate how well a generated reasoning answer matches the ground truth reasoning for a given question.

You will evaluate the reasoning answer on a scale of 1-5. 5 means the generated reasoning accurately reflects the same facts, logic, and overall conclusion as the ground truth reasoning. 1 means the generated reasoning presents contradictory facts, logic, or reaches an opposite conclusion compared to the ground truth reasoning.

Consider both the direct answers and reasoning answers provided when evaluating the reasoning. Crucially, if the generated_direct_answer

118

fundamentally contradicts the `ground_direct_answer` (e.g., 'Yes' vs. 'No', or stating an object is present when it's absent), then the `generated_reasoning` is supporting an incorrect conclusion. In such cases, even if the `generated_reasoning` discusses similar elements or topics as the `ground_reasoning`, it cannot be considered a good match and the `reasoning_score` must be low (typically 1, or 2 if there's any marginal, non-contradictory similarity in how the reasoning is framed despite the factual error).

Example:

Question: What safety hazards are visible in the warehouse?

Ground truth direct answer: Exposed cables and scattered materials

Generated direct answer: Cables on the floor

Ground truth reasoning: The video shows exposed cables crossing walkways and packaging materials scattered on the floor creating trip hazards.

Generated reasoning: There are cables running across the floor that could cause workers to trip.

Output:

```
{  
  "reasoning_score": 3  
}
```

Your Turn:

Question: {question}

Ground truth direct answer: {ground_direct_answer}

Generated direct answer: {generated_direct_answer}

Ground truth reasoning: {ground_reasoning_answer}

Generated reasoning: {generated_reasoning_answer}

Output JSON Format:

```
{  
  "reasoning_score":  
}
```

References

- [1] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. [8](#)
- [2] M. Khanov, J. Burapachee, and Y. Li. ARGS: Alignment as reward-guided search. In *The Twelfth International Conference on Learning Representations*, 2024. [8](#)
- [3] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. [8](#)
- [4] G. Li, H. Du, and D. Hu. Boosting audio visual question answering via key semantic-aware cues, 2024. [8](#)
- [5] W. Li, X. Zhang, S. Zhao, Y. Zhang, J. Li, L. Zhang, and J. Zhang. Q-insight: Understanding image quality via visual reinforcement learning, 2025. [8](#)
- [6] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let’s verify step by step, 2023. [8](#)
- [7] L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun, and A. Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024. [8](#)
- [8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. [8](#)
- [9] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. [8](#)
- [10] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. [8](#)
- [11] P. Wang, L. Li, Z. Shao, R. X. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024. [8](#)
- [12] Y. Wang, Z. Li, Y. Zang, C. Wang, Q. Lu, C. Jin, and J. Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning, 2025. [8](#)
- [13] Y. Wang, Z. Tan, J. Wang, X. Yang, C. Jin, and H. Li. Lift: Leveraging human feedback for text-to-video model alignment, 2025. [8](#)
- [14] Y. Wang, Y. Zang, H. Li, C. Jin, and J. Wang. Unified reward model for multimodal understanding and generation, 2025. [8](#)
- [15] T. Xiong, X. Wang, D. Guo, Q. Ye, H. Fan, Q. Gu, H. Huang, and C. Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. [8](#)
- [16] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 3480–3491, New York, NY, USA, 2022. Association for Computing Machinery. [8](#)