EgoAVU: Egocentric Audio-Visual Understanding

Supplementary Material

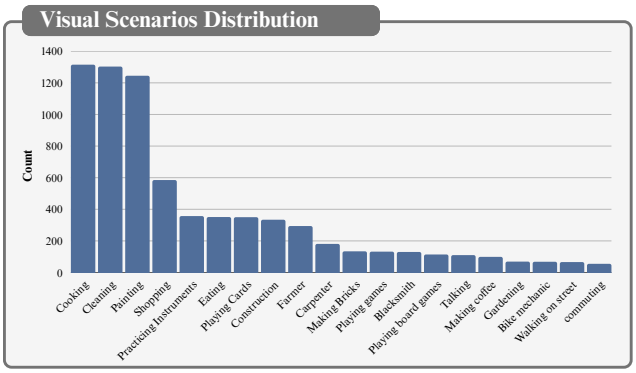

Figure 6. **Distribution of 20 most common visual scenarios in EgoAVU-*Instruct* and EgoAVU-*Bench***

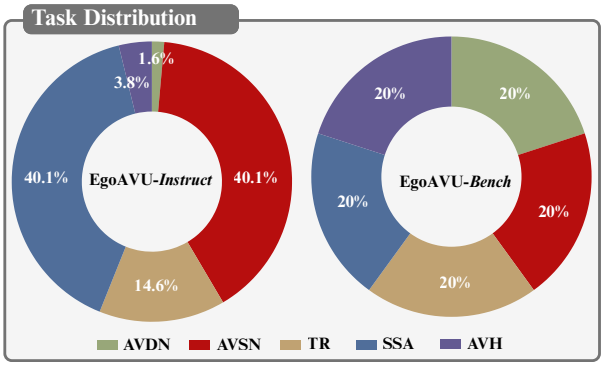

Figure 7. **Distribution of proposed tasks across EgoAVU-*Instruct* and EgoAVU-*Bench***

A. Additional Details on EgoAVU

A.1. Prompts

Prompt Used. We describe the various prompts used in EgoAVU. These prompts cover the generation of multiple modules including the Multi-modal Context Graph (see Fig. 11), Audio-Visual Narration (see Fig.13), and task-specific QA generation (see Fig. 14, Fig. 15, Fig. 16, Fig. 17, Fig. 18, Fig. 19, Fig. 20). The general structure of each prompt includes: (i) defining the objective, (ii) specifying the input format, (iii) describing the task, (4) providing general instructions, (iv) including human-generated examples to enable in-context learning, and (v) specifying the output format.

A.2. MATTR

For video filtering, we utilize the Moving-Average Type-Token Ratio (MATTR) to identify videos with rich multi-modal diversity. We set the window size to $w = 200$ tokens based on the average token length of combined narrations across video segments in our dataset, ensuring the window captures sufficient multi-modal context for meaningful diversity measurement. We retain videos whose MATTR exceeds $\tau = 0.3$, effectively removing the bottom 25% of our distribution to filter out static or repetitive descriptions. This threshold was determined through manual inspection of 100 randomly sampled videos across different MATTR ranges. Videos below $\tau = 0.3$ predominantly featured repetitive actions with limited object diversity and minimal auditory variation, while videos above this threshold exhibited richer multimodal dynamics (refer to Fig 10 for more examples).

A.3. Ablation on Multi-Modal Context Graphs

To validate the necessity of the Multi-Modal Context Graphs (MCG) component in EgoAVU, we conducted an ablation experiment on 200 randomly sampled video clips. We compared our MCG-based pipeline against a direct baseline where LLaMA-3-70B generates audio-visual narrations directly from enhanced narrations (video caption, image caption, audio caption, and action narration) without the intermediate MCG structure. We manually evaluated both output, assessing: (1) completeness of sound-source associations, (2) accuracy of action sequences, and (3) overall audio-visual coherence. We observed that the direct method produced errors in 82 out of 200 captions (41.0%), with the breakdown as follows: 48 captions (19.0%) missed or incorrectly associated sound sources, 31 captions (15.5%) omitted crucial action sequences or interaction details, and 17 captions (3.5%) exhibited both issues (refer to Fig. 8 for example). In contrast, the our MCG-based approach reduced errors to 21 out of 200 captions (10.5%), representing a 76.1% relative error reduction.

B. Manual Effort for EgoAVU-*Bench* Construction

To ensure the reliability of EgoAVU-*Bench*, we conducted extensive manual verification across all 3,000 question-answer pairs covering 900 egocentric videos. Each video was carefully reviewed by trained annotators, taking approximately 2-3 minutes per video to verify temporal alignment and audio-visual correspondence. Out of 3,000 QA pairs, 1,524 pairs (50.8%) were modified or corrected during this process. For open-ended tasks (SSA, AVDN,

Figure 8. **Qualitative comparison of audio-visual narrations generated with and without (w/o) Multi-modal Context Graph (MCG).** Our MCG-based approach produces narrations with superior audio-visual coherence, accurately capturing action sequences and sound-source associations, while the direct method (w/o MCG) often misses critical sounds or action sequences.

AVSN), corrections primarily addressed missing sounds, incorrect human-object interactions, and sound-source misalignments to ensure accurate audio-visual grounding. For close-ended tasks (TR, AVH), we verified answer correctness and enhanced distractor quality by ensuring multiple-choice options were sufficiently challenging and plausible while avoiding options that were too similar to correct answers or obviously incorrect. The complete manual verification process required approximately 225 hours of human annotation effort.

C. Additional Details on Narration Enhancement

Prompts. To capture spatial details, we extract the center frame and prompt Qwen2.5-VL with ***"Identify all the objects visible in the image"*** to detail all objects present in the video clip. To capture temporal dynamics, we utilize Qwen2.5-Omni in two stages: first, we process only the video frames with the prompt ***"Describe the video in detail"*** to capture visible activities; then, we process the audio with the prompt ***"Describe all the sounds heard in detail"*** to capture auditory information.

D. Data statistics

In this section, we provide additional data statistics. As shown in Fig. 6, both EgoAVU-*Instruct* and EgoAVU-*Bench* encompass egocentric videos spanning diverse visual scenarios, including cooking, cleaning, and outdoor/indoor activities. Furthermore, Fig. 7 illustrates the distribution of the five proposed tasks across EgoAVU-*Instruct* and EgoAVU-*Bench*.

E. Additional Experiment Details

Training Details. For both LoRA and full fine-tuning, we use a maximum context length of 30,000 tokens and sample videos at 1 FPS with a frame resolution of 256×256. Training is performed using DeepSpeed ZeRO-3 with a learning rate of 1×10^{-4} and a cosine schedule with 10% warmup over 5 epochs. We perform balanced sampling, i.e, sample with equal weights from each task, during training.

Evaluation Details. Fig. 21 presents the prompt used for our *LLM-as-judge* evaluation. Following prior work [30], we assess the reliability of LLM-based scoring by measuring its alignment with human judgments on 300 randomly sampled open-ended QA pairs from EgoAVU-*Bench*. The resulting human-alignment rate is 87.6%, indicating strong

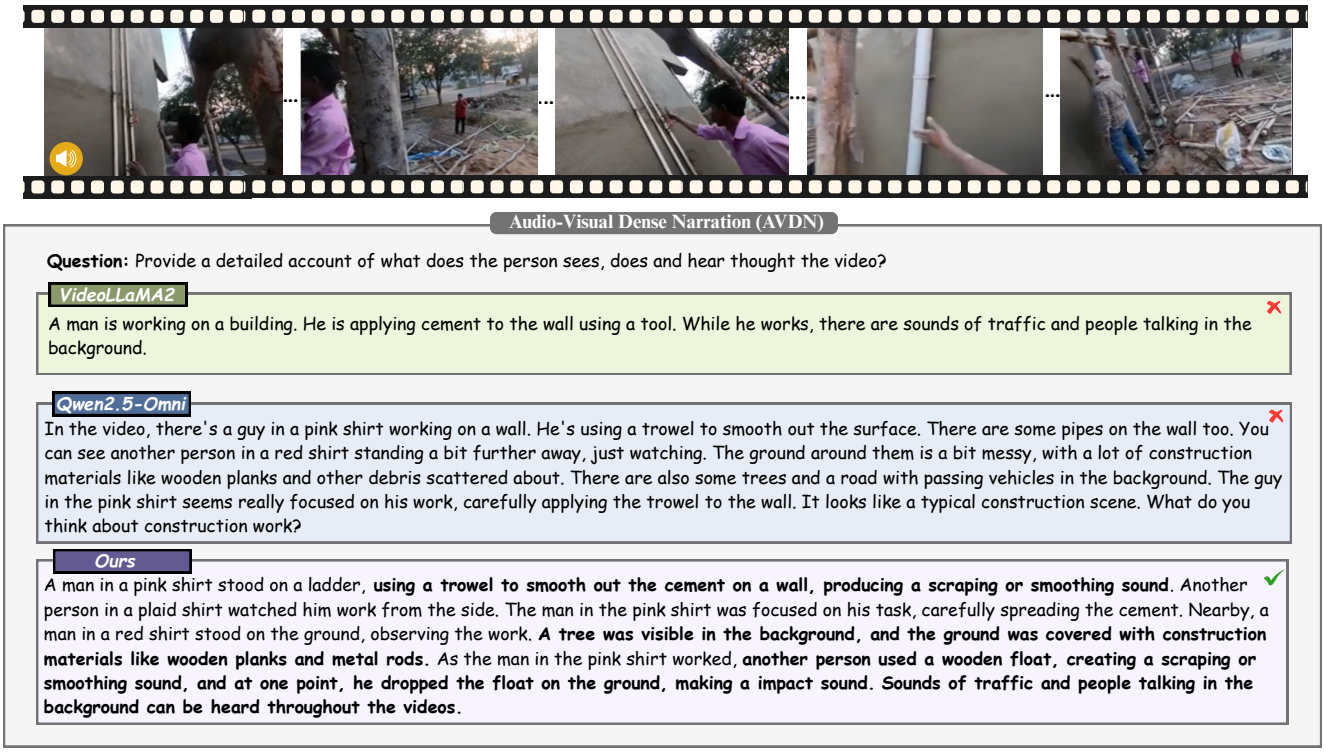

Figure 9. **Qualitative comparison of various MLLMs on the Audio-Visual Dense Narration (AVDN) task.** Our model fine-tuned on EgoAVU-*Instruct* captures significantly more dense visual details than Qwen2.5 Omni and VideoLLaMA2, while also identifying auditory cues related to human actions and background sounds in the video.

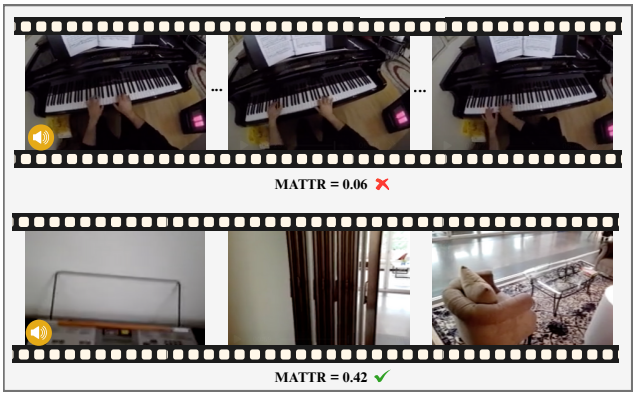

Figure 10. **Examples of video filtering based on MATTR scores.** Videos with low MATTR (0.06) show repetitive content (person playing piano throughout the video.), while high MATTR (0.42) indicates rich multimodal diversity across human actions, visible objects, and ambient sounds.

alignment between the two.

F. Additional Results

In addition to egocentric benchmarks, we evaluate our fine-tuned model on popular exocentric Video QA benchmarks, including VideoMME (Short duration split w/o subtiles) and AVQA. As shown in Table 7, despite being fine-tuned exclusively on egocentric QAs, our model almost retains its original performance on VideoMME and slightly outperforms the base model Qwen2.5-Omni on audio-visual QAs in AVQA.

Model	VideoMME Acc. (↑)	AVQA Acc. (↑)
Qwen2.5-Omni	**73.0**	89.4
OURS $_{(LoRA)}$	72.4	**89.7**
OURS $_{(Full)}$	72.0	89.5

Table 7. **Result on other Exocentric Video QA benchmarks.**.

.

G. Qualitative Analysis

We qualitatively validate our pipeline and model performance. Fig. 8 shows that our *Multi-Modal Context Graph (MCG)* captures key sounds and actions that direct narration often misses (e.g., “clattering”). Fig. 9 highlights that, unlike our fine-tuned model, existing MLLMs produce sparse descriptions for Audio-Visual Dense Narration task

in EgoAVU-*Bench* and further overlook audio cues or fail to ground sounds to their sources. Fig. 10 illustrates our *MATTR-based filtering*: low-scoring videos (0.06) contain repetitive content, whereas high-scoring ones (0.42) exhibit diverse multimodal activity.

Objective: You are an AI assistant tasked with performing a high-fidelity analysis of video content. Your role is to function as an evidence extractor, not an open-world reasoner. You must strictly use the provided captions to identify object interactions and to analyze sounds, grounding every piece of information directly to the source text.
Inputs: You will be provided with a JSON object containing the following four keys:
- *Video Caption:* Describes the overall visual scene, including events and actions of entities other than the narrator (e.g., animals, other people, environmental events).
- *Image Caption:* Describes the diverse objects visible in the center frame.
- *Audio Caption:* A transcript of sounds and audio events.
- *Action Narration:* Describes the specific actions performed by the primary person ($\#CC$ denotes person)

Task: Analyze the provided captions to generate a structured multimodal context graph in the form of JSON that captures multimodal relationship. For generating the multimodal context graph, follow the instruction mentioned below:
Instructions:
- *Identify Interacted Objects:* Parse the action narration to find all objects the primary person is described as touching, holding, using, or manipulating. Compile these into the "interacted objects" list with what action the person performed.
- *Identify Background Objects:* Take the complete list of objects from image caption. Create a new list containing only the items from image caption that are NOT in your "interacted objects" list. This will be your "non interacted objects" list.
- *Identify Sound-Source Associations* This is a strict, evidence-based process. Ideally there can be two type of sound, sound caused by action/object or background sound. Your task is to capture both of them for audio caption
  A. Find the Grounding Evidence for foreground sound: For each sound in audio caption, you must search action narration and video caption to find the specific text that describes the action or event causing it. Look in action narration for causes related to the primary person's actions. Look in video caption for causes related to other entities (animals, other people) or general scene events. ambient sound include music, background noise, etc. Crucially: If no direct textual evidence can be found in either caption that explains the sound's origin, the sound is a background sound.
  B. Exclude "Unquestionable" Sounds: Even if grounded, do NOT include a sound if it falls into these categories: Mundane Biological Sounds: Common sounds like "breathing," "sighing," "swallowing." Vague Ambient Noise: like "white noise" or "faint hum."
  C. Determine the sound category: Classify as Foreground Sound (from the human action or visible object) or Background Sound.
  D. Handle Empty Results: If no sounds pass the filtering and grounding process, the "sounds" list in your output must be an empty list ([]).

Human Generated Examples: Here are 5 human created examples for the correct execution <examples>.
Important Note:
- In the example how "giggle" was excluded because it had no grounding evidence in video caption or action narration.
- Source description will contain either the corresponding action or the object that can produced the sound
- Sound category can have foreground sound such as ''Action Sound'', ''Object Sound'' or background such as ''Ambient Sound''

Final Output Format: Your entire response MUST be a single, valid JSON object following the structure of the example. Do not include any text outside of the JSON structure.
Here is the input: <input>

Figure 11. **Prompt For Generating Multi-Modal Context Graphs**

```
{
        "interacted_objects": [
            ["sink", "#C C rinses both hands"],
            ["tap", "#C C turns on tap"],
            ["door", "#C C opens the door"],
        ],
        "background_objects": [
            "oranges",
            "sponge",
            "red chair",
            "microwave",
            "cabinets"
        ],
        "sounds": [
            {
                "acoustic_description": "water flowing sound",
                "source": "#C C turns on tap",
                "evidence_source": "action_narration",
                "sound_category": "Foreground Sound"
            },
            {
                "acoustic_description": "hands being rinsed sound",
                "source": "#C C rinses both hands",
                "evidence_source": "action_narration",
                "sound_category": "Foreground Sound"
            },
            {
                "acoustic_description": "door opening and closing sound",
                "source": "#C C opens the door",
                "evidence_source": "action_narration",
                "sound_category": "Foreground Sound"
            }
        ]
    }
```


Figure 12. **Example of MCG.** Example of MCG generated in JSON format using the above-mentioned prompt.

Objective: Your job is to generate a single, detailed, and objective paragraph summarizing what can be seen and heard in the video clip.

Input: You will be given:

- A free-form natural language paragraph summarizing what happens in the video. This is typically derived from a loose transcription or human description of the scene.
- A list of short, possibly overlapping or action tags extracted from the video. These may contain minor inconsistencies, but offer clues about human interactions and movements in the scene. Focus closely on the interaction starting with #C C.
- A multi-modal context graph represented as a structured JSON object containing: "interacted_objects" --- objects the person interacted with, "non_interacted_objects" --- objects present but not interacted with, and "sounds" --- sound events grounded in the narration, with descriptions and causes.

Task: Write a single, coherent paragraph that summarizes the video scene in detail, following these rules:

Instructions

- Clearly describe all key actions in the scene, combining the raw description and narration tags.
- Include all interacted objects, and describe how each was interacted with.
- Mention all non-interacted objects that are visible or relevant once, integrated naturally into the scene description. Do not repeat them again at the end.
- Integrate sound events by describing what caused them and when, grounded in the referenced actions.
  - Not all actions have corresponding sounds.
  - Only include sounds that are listed in the scene graph.
  - Use semantically appropriate or naturalistic descriptions for acoustic events. For instance, if the sound is caused by shaking a spray bottle, you may refer to it as ``crunching or rattling''.
- Use an objective and factual tone|avoid any emotional, subjective, or evaluative language (e.g., no ``cute,'' ``interesting,'' or ``simple'').
- Write in past tense.
- Ensure the paragraph flows naturally and avoids redundancy.

Human Generated Examples: Here are 5 human created examples for the correct execution <examples>.

Final Output Format: The final output must be in JSON format with key as "caption"

Here is the input to generate the caption: <input>

Figure 13. **Prompt For Generating Audio-Visual Narration**

Objective: You are an AI assistant tasked with analyzing a video segment and performing three tasks:

- Generate a single open-ended question about the sound-source association observed in the video.
- Produce a natural, human-like narration that links sounds to the actions and objects responsible.
- Generate a detailed, structured answer to the question, grounded entirely in the provided scene graph metadata.

Input: You will be provided with:

- video description: description of the video segment
- Multi-modal Context Graph: <Details on Multi-modal Context Graph>

Instructions: Follow the below instruction to complete the task:

- *Question Generation.* If the "sounds" list is empty or missing, return this exact string as the only output: "No significant sound is present in the video clip." Otherwise, use the template below: <template>
- When narrating egocentric data, a person is sometimes referred to by capital letters, such as "C." When writing the description, treat such IDs as referring to a person. For example, if a sound-producing evidence states "person C is clapping," it should be treated as "the person is clapping."
- *Detailed Answer Generation.*
  - Structure the answer as follows: Begin with a sentence that clearly states how many distinct grounded sound events were present. Then provide one sentence for each sound, explaining what caused it by using the acoustic description and grounding evidence. Treat C as the person in the video.
  - Do not speculate or add interpretation beyond the metadata.
  - Do not include any text outside of the JSON structure.
  - Do not include any step by step explanations.

Human Generated Examples: Here are 5 human created examples for the correct execution <examples>.

Output Format: Output must be in a JSON format with following key "question" and "answer".

Here is the input to generate the question-answer pair: <input>

Figure 14. **Prompt for generating Sound-Source Association Question-Answer pair**

Objective: Generate two sound-related question--answer pairs from an egocentric video caption that describes a person's visible actions, sounds, and objects. The output should be formatted as JSON with one correct and one hallucinated sound question.
Input: You will be given an egocentric video narration containing descriptions of:
- The person's visible actions
- Distinctive sounds (e.g., hissing, tapping, scraping)
- Objects present in the scene
- Temporal information about when events occur

Instructions: Follow the instruction below:
- *Focus on distinctive sounds* such as foreground sounds related to human-object interaction such as hissing, tapping etc. or background sounds such as bird chirping etc.
- *Generate one correct question:* Ask about a sound explicitly mentioned in the narration.
- *Generate one hallucinated question*: Ask about a plausible sound that is *not* mentioned in the narration.
- *Answer format*: Answers must be in binary format "Yes" or "No".

Output Format: The output must be in JSON format with following keys: "question", "question type" including "Factual", "Hallucinated" and "answers".
Here is the input to generate the question-answer pair: <input>

Figure 15. **Prompt for generating Audio-Visual Hallucination (Sound) Question-Answer pair**

Objective: Generate two action-related question--answer pairs from an egocentric video caption that describes a person's visible actions, sounds, and objects. The output should be formatted as JSON with one correct and one hallucinated action question.
Input: You will be given an egocentric video narration containing descriptions of:
- The person's visible actions
- Distinctive sounds (e.g., hissing, tapping, scraping)
- Objects present in the scene
- Temporal information about when events occur

Instructions: Follow the instruction below:
- *Focus on distinctive, non-trivial actions* such as wiping, twisting, or squeezing. Avoid trivial actions such as breathing, walking, or placing.
- *Generate one correct question:* Ask about an action explicitly mentioned in the narration.
- *Generate one hallucinated question*: Ask about a plausible action that is *not* mentioned in the narration.
- *Answer format*: Answers must be in binary format "Yes" or "No".

Output Format: The output must be in JSON format with following keys: "question", "question type" including "Factual", "Hallucinated" and "answers".
Here is the input to generate the question-answer pair: <input>

Figure 16. **Prompt for generating Audio-Visual Hallucination (Action) Question-Answer pairs**

Objective: Generate two object-related question--answer pairs from an egocentric video caption that describes a person's visible actions, sounds, and objects. The output should be formatted as JSON with one correct and one hallucinated object question.
Input: You will be given an egocentric video narration containing descriptions of:
• The person's visible actions
• Distinctive sounds (e.g., hissing, tapping, scraping)
• Objects present in the scene
• Temporal information about when events occur
Instructions: Follow the instruction below:
• *Focus on specific, manipulable objects.* Avoid generic nouns like 'things', 'stuff', or 'material'.
• *Generate one correct question:* Ask about an object explicitly mentioned in the narration.
• *Generate one hallucinated question*: Ask about a plausible object that is *not* mentioned in the narration.
• *Answer format*: Answers must be in binary format "Yes" or "No".
Output Format: The output must be in JSON format with following keys: "question", "question type" including "Factual", "Hallucinated" and "answers".
Here is the input to generate the question-answer pair: <input>

Figure 17. **Prompt for generating Audio-Visual Hallucination (Object) Question-Answer pairs**

Objective: Generate two temporal reasoning question--answer pairs from a list of chronological video narrations, focusing on the order of Action, Object, and Sound events. The output should be formatted as a JSON list containing one "before" and one "after" question.
Input:
• A list of narration describing what happens in the video in chronological order ({caption_list}).
• The specific question type to be generated ({type}: one of Action-Action, Action-Object, or Action-Sound).
Instructions: Follow the steps below:
1. *Identify Distinct Events:* Identify several unique, non-trivial, and non-repetitive events, each describing an **Action**, an **Object**, or a **Sound**.
2. *Select Event Pair (E1, E2):* Choose two events occurring at different times that match the required category ({type}). E1 must chronologically precede E2.
3. *Generate Questions:* Create one "before" question (referencing E2) and one "after" question (referencing E1) using the corresponding template:
   - **Action--Action**
     - Before: "What action was the person performing before <E2>?"
     - After: "What action did the person perform after <E1>?"
   - **Action--Object**
     - Before: "What objects can be seen before the person performed the <E2> action?"
     - After: "What objects can be seen after the person performed the <E1> action?"
   - **Action--Sound**
     - Before: "What sound can be heard before the person <E2>?"
     - After: "What sound can be heard after the person <E1>?"
4. *Answer and options:* Write a concise, naturalistic answer as if you watched the video. Include three plausible options that fit the context but are temporally incorrect.
Output Format: The output must be a JSON list of exactly two question objects (one "before" and one "after") with the following keys: "question", "answer", "type", and "options".

Figure 18. **Prompt for generating Temporal Reasoning (before/after) Question-Answer Pairs**

Objective: Generate one multiple-choice question about the temporal order of four events derived from a sequence of chronological egocentric video narration.
Input: A sequence of detailed narration in chronological order describing what happens in a egocentric video.
Instructions: Follow the steps below:
- *Identify Four Grounded Events:* Identify four unique, non-trivial events. Each event must include concise but meaningful details about the person's activity, the visible surroundings, and any sounds mentioned.
- *Grounding Constraint:* All events must be directly derived from the narrations. Do not hallucinate or invent any objects, sounds, or actions.
- *Create Temporal Question:* Create one general multiple-choice question that asks about the temporal order of the four events (e.g., "Which event happened first?", "Which moment occurred last?").
- *Create Options:* List the four events as options A, B, C, and D. Ensure the description for each option is the exact description provided in the events list.
- *Provide Correct Answer:* Indicate the correct temporal order by selecting one of the options (A, B, C, or D) as the correct_answer.

Output Format: The output must be in JSON format with the following keys: "events" (a list of the four descriptions), "question", "options" (a map of A, B, C, D to the event descriptions), and "answer".

Figure 19. **Prompt for generating Temporal Reasoning (Event Ordering) Question-Answer Pairs**

Objective: Write a single, coherent, dense narration summarizing the entire video based on a list of 10-second captions.
Input: A list of narration, each describing a 10-second segment of a egocentric video, including start_time, end_time, and the caption text.
Instructions: The final output must be a single, fluent paragraph that acts as a dense narration. The paragraph must adhere to the following rules:
- Integrate all actions, objects, and sounds across the full video.
- Use timestamps in seconds to indicate when key events occurred.
- Group similar or adjacent events into continuous spans.
- Avoid listing or repeating captions verbatim.
- Use only the information in the input captions.
- Be concise and fluent.
- Do not invent any new information or context.

Human Generated Examples: Here are 3 human created dense narration: <examples>
Output Format: A single paragraph, not a JSON object.
Here is the input: <input>

Figure 20. **Prompt for generating Audio Visual Dense Narration**

Objective: Act as an impartial grader to evaluate a PREDICTED_ANSWER against a GROUNDING_ANSWER with respect to a QUESTION.

Input:

- QUESTION: The question posed to the model.
- GROUNDING_ANSWER: The authoritative reference answer.
- PREDICTED_ANSWER: The model's answer to be graded.

Instructions for Grading:

1. *Comparison:* Compare PREDICTED_ANSWER to GROUNDING_ANSWER with respect to the QUESTION.
2. *Assign Rating (1–5, integer only):*
   - **5:** Fully correct, complete, and faithful to the grounding; no meaningful errors or omissions.
   - **4:** Mostly correct; minor omissions or small inaccuracies that do not change the overall correctness.
   - **3:** Partially correct; captures some key points but misses important details or includes notable inaccuracies.
   - **2:** Largely incorrect; substantial errors, contradictions, or missing major required points.
   - **1:** Incorrect/irrelevant; contradicts the grounding or fails to answer the question.
3. *Provide Reasoning:* Briefly explain the rating (1{4 concise sentences).

Judging Rules (Priorities):

- Prioritize factual alignment with the GROUNDING_ANSWER. Contradictions result in heavy penalization.
- Extra details are acceptable only if they do not conflict with the grounding and remain relevant to the QUESTION.
- Penalize hallucinations, unverifiable claims, safety issues, and failure to address the core of the QUESTION.
- Do not reward verbosity or style unless it improves factual accuracy or completeness with respect to the grounding.
- If the grounding indicates the question is unanswerable, judge whether the prediction correctly reflects that.

Output Format: The output **MUST** be valid JSON (no markdown, no extra text) with the following keys:

```
{
  "rating": <int value between 1 to 5>,
  "reason": "<string of 1-2 lines explaining the rating>"
}
```

Figure 21. **Prompt for LLM-as-judge evaluation**