HAVE-Bench: Hierarchical Audio-Visual Evaluation from Perception to Interaction

Supplementary Material

7. Ethical, Legal, and Data Release Statement

All external media used in this work are obtained from publicly accessible sources under appropriate licenses. Specifically, all Street View imagery is accessed through the Google Maps API under the Google Maps Platform Terms of Service. To comply with Google's content redistribution policy, we do not redistribute the raw imagery. Instead, we release only metadata (e.g., latitude, longitude, heading, pitch), which allows researchers to reproduce the identical observations via the official API.

To ensure reproducibility while maintaining full legal and ethical compliance, we will release all benchmark data, annotations, and evaluation tools upon acceptance of the paper. Components covered by third-party usage restrictions will be distributed in metadata-only form, enabling faithful reconstruction without violating any licensing terms.

8. Evaluation Details

8.1. Human Evaluation

As described in Sec. 3.6, HAVE-Bench relies on GPT-4o as the LLM judge for multiple-choice, open-ended, and interaction-level tasks. To assess the reliability of GPT-4o's judgments, we conduct a stratified human evaluation to measure its alignment with expert annotators.

Evaluation Setup. We randomly sample 150 items from each of the three evaluation types in HAVE-Bench (multiple-choice, open-ended, and interaction), yielding 450 evaluation items. For each item, we obtain answers from three representative systems— the strongest closed-source model (Gemini2.5-Flash), the strongest open-source model (Qwen2.5-Omni), and a mid-tier open-source model (MiniCPM-o). Thus, human annotators assess a total of **1350 model–item judgments** (450 items $\times$ 3 models).

Annotators are shown the image, audio (or its transcript when necessary), the question, and the model's answer, and must label the answer as *correct* or *incorrect* following the same rubric used by the GPT-4o judge. Specifically, for interaction tasks—which are closed-loop and therefore cannot be reliably evaluated at the episode level once an early action alters the trajectory—we follow the same step-wise protocol as the GPT-4o judge: annotators determine whether each predicted action satisfies the success criterion for the current state.

Each model–item judgment is independently labeled by two annotators. We compute (i) raw human–human agreement, (ii) raw human–GPT-4o agreement, and (iii) Cohen's κ for the human–GPT-4o pair, where κ measures chance-corrected consistency.

Table 3. **Human–Human and Human–GPT-4o agreement on 1350 judgments** (from 450 items and 3 models). Agreement is reported as a percentage. Cohen's κ reflects chance-corrected consistency.

Eval. Type	Human–Human (%)	Human–GPT (%)	κ
Multiple-choice	98.5	98.1	0.95
Open-ended	98.0	96.5	0.93
Interaction	97.3	96.8	0.93
Overall	**97.9**	**97.1**	**0.94**

Results. As shown in Table 3, GPT-4o exhibits high agreement with human annotators across all evaluation types (97.1% overall), with strong chance-corrected consistency ($\kappa = 0.94$). Even for interaction tasks—which involve multi-step reasoning and action prediction— the human–GPT agreement reaches 96.8%. The human–human agreement (97.9%) is only slightly higher than the human–GPT agreement, indicating that GPT-4o approximates expert judgment closely. Most discrepancies arise from borderline partially-correct responses.

Importantly, the interaction-level agreement is computed at the *step level*, where annotators judge whether each predicted action satisfies an explicit state-transition success criterion. The high consistency (96.8%) therefore reflects the clarity of our state annotations and transition rules, and confirms that interaction evaluation in HAVE-Bench is well-defined and not sensitive to subjective interpretation. Overall, these findings demonstrate that GPT-4o provides a reliable and scalable evaluation signal for HAVE-Bench, with judgment quality comparable to human assessment.

8.2. Confidence Intervals and Variance Analysis for Interaction Tasks

Interaction-level subtasks in HAVE-Bench are relatively small (Navigation: 87 items, Music: 100 items, Puzzle: 76 items), and model accuracies on these tasks are generally low (mostly 5–30%; Table 1). Under such conditions, point estimates alone may not reliably indicate true model performance. We report (i) 95% binomial Wilson confidence intervals, which capture statistical uncertainty arising from finite sample sizes, and (ii) standard deviations over three

evaluation runs with different random seeds, which captures model-side variability independent of dataset size.

The results are shown in Table 4. Despite noticeable sampling uncertainty (Wilson CI widths of ± 5–10 points) and modest run-to-run variation ($\sigma < 5$ points for all models), the closed–open performance gaps reported in the main results remain substantially larger than either source of variance (e.g., 30.4 vs. 18.0 on Interaction-Avg.). This indicates that the relative ranking of models and the overall difficulty of The interaction level is robust to statistical fluctuations, and that our Conclusions do not hinge on individual samples or random-seed effects.

8.3. Evaluation Prompt

We use GPT-4o as a unified LLM judge across all evaluation types. For multiple-choice questions, we adopt the MMMU-style [60] judging prompt from VLMEvalKit [11], which matches the model's free-form answer to the most semantically consistent option. For open-ended questions, we use a correctness-judging prompt (shown in Figure 4) that compares the model's answer with the reference under semantic equivalence and factual-consistency criteria. Interaction tasks use the same open-ended judging prompt, except that the ground truth is the natural-language success criterion for The current state transition; the judge determines whether the predicted action satisfies this criterion.

9. Additional Ablations on SI vs. TI Modes

In Sec. 4.4 (Table 2), we showed a consistent gap between text–image (TI) and speech–image (SI) modes on Audio-as-Instruction tasks. We hypothesize that this gap is mainly due to the model's failure to fully transfer its text–image reasoning capability to speech–image joint reasoning, rather than artifacts introduced by our data pipeline. To support this claim, we conduct several ablations designed to rule out two potential confounding factors: (i) the spoken-form rewriting step used to make questions TTS-friendly, and (ii) the choice of speech source (TTS vs. human-read).

Spoken-Form vs. Original Text We first evaluate whether our spoken-form rewriting process alters task difficulty. For each Audio-as-Instruction (AaI) sample, we compare: (i) **TI-raw**: the original question text from the VQA dataset; (ii) **TI-spoken**: the rewritten spoken-form text generated during the TTS pipeline; and (iii) **SI**: the speech–image mode that serves as a baseline. This experiment isolates whether the TI–SI gap could be attributed to artifacts introduced during the spoken-form rewriting step rather than to the speech modality itself.

As shown in Table 5, TI-raw and TI-spoken achieve nearly identical accuracy across all models ($|\Delta| <$ 1–2 points), while both are clearly above SI. This confirms that spoken-form rewriting does not change task difficulty in a systematic way, and rules out the rewriting step as the main cause of the TI–SI performance gap.

TTS vs. Human-Read Speech Next, we examine whether the TI–SI gap is influenced by the quality or style of the spoken audio. We compare three input settings for the same AaI samples: (i) **SI-Human**, using professionally recorded speech for a subset of 400 samples; (ii) **SI-TTS**, using Azure neural TTS for the remaining samples; and (iii) **TI**, the text–image mode that serves as a baseline. This experiment evaluates whether differences in speech source (human vs. TTS) affect speech–image comprehension and whether such variation could partially explain the observed TI–SI performance gap.

Table 6 reports SI accuracy under TTS and human-read speech, with TI accuracy shown as a text-based reference. Across all models, SI-TTS and SI-Human mostly differ by 1–3 points and neither source consistently outperforms the other. These inconsistent differences indicate that the TI–SI gap cannot be explained by artifacts specific to TTS synthesis or by a strong preference for human-read speech, further ruling out the speech source as the primary driver of the observed performance difference.

10. Data Annotation Details

This section details the annotation procedures used to construct HAVE-Bench across three task families: Audio-as-Instruction, Audio-as-Context, and interaction-level tasks. We describe how written questions are converted into natural spoken form, how GPT-4o-assisted scoring and downsampling are applied for quality control, how template-based question design guides AaC annotations, and how multi-turn task graphs are defined for interactive scenarios.

10.1. Audio-as-Instruction Tasks

Prompts of spoken form conversion Table 7 presents the prompts used to convert written Q&A questions into natural spoken form using GPT-4o. The goal of this transformation is to ensure that LaTeX expressions, mathematical symbols, and structural placeholders are converted into fluent, intelligible speech suitable for text-to-speech (TTS) synthesis and robust multimodal evaluation. Examples of original vs. spoken-form questions are shown in Figure 5.

GPT-4o downsampling As described in Sec. 3.5, we apply a quality-based downsampling procedure to ensure balanced and high-quality coverage across the AaI sub-tasks. Each candidate sample is evaluated by GPT-4o using a unified multimodal–QA scoring rubric, which rates accuracy, logical coherence, clarity, relevance, and engagement/depth on a 0–10 scale. Low-quality or redundant items are filtered out, and sampling weights are adjusted to preserve diversity

Table 4. **Statistical uncertainty on interaction-level tasks.** We report mean accuracy (Acc, %), standard deviation across three runs (σ, in percentage points), and 95% Wilson confidence intervals (CI, %).

Model	Navigation(87)			Music(100)			Puzzle(76)		
	Acc	σ	CI	Acc	σ	CI	Acc	σ	CI
Gemini2.5-Flash	22.8	3.0	[15.4, 32.9]	37.7	2.7	[29.1, 47.8]	30.7	2.3	[21.1, 41.3]
Qwen2.5-Omni	14.2	2.2	[8.1, 22.6]	19.2	1.7	[12.5, 27.8]	20.6	3.5	[13.4, 31.5]
MiniCPM-o	12.3	4.0	[7.2, 21.2]	14.2	1.1	[8.5, 22.1]	15.0	1.3	[8.3, 24.1]
Ming-Lite-Omni	8.0	4.0	[4.0, 15.7]	15.8	1.1	[10.1, 24.4]	11.0	1.3	[5.4, 19.4]
VITA-1.5	3.1	1.1	[1.1, 9.7]	7.3	2.2	[3.4, 13.7]	5.7	3.5	[2.1, 12.8]
Megrez-3B-Omni	2.5	1.1	[0.6, 8.0]	7.7	2.3	[4.1, 15.0]	1.3	0.0	[0.2, 7.0]

Table 5. **Comparison between original text (TI-raw) and spoken-form text (TI-spoken).** SI is provided as a reference. Δ denotes TI-raw $-$ TI-spoken.

Model	SI	TI-raw	TI-spoken	Δ
Gemini2.5-Flash	71.3	73.2	72.7	0.5
Qwen2.5-Omni	69.1	70.9	71.5	-0.6
MiniCPM-o	58.1	61.1	62.1	-1.0
Ming-Lite-Omni	65.7	71.3	70.6	0.7
Ola	62.6	68.0	67.8	0.2
VITA-1.5	61.8	64.9	63.4	1.5
Megrez-3B-Omni	47.8	59.3	59.0	0.3

Table 6. **TI baseline vs. SI performance under TTS and human-read speech.** Δ denotes SI-TTS $-$ SI-Human.

Model	TI	SI-TTS	SI-Human	Δ
Gemini2.5-Flash	73.2	71.0	72.2	-1.2
Qwen2.5-Omni	70.9	69.7	68.0	$+1.7$
MiniCPM-o	61.1	58.4	57.6	$+0.8$
Ming-Lite-Omni	71.3	65.8	65.5	$+0.3$
Ola	68.0	62.1	63.9	-1.8
VITA-1.5	64.9	62.0	61.3	$+0.7$
Megrez-3B-Omni	59.3	48.5	45.9	$+2.6$

in domain, structure, and difficulty. We provide the exact prompt in Figure 6.

10.2. Audio-as-Context Tasks

Question Templates of Audio-as-Context Tasks To standardize question design for Audio-as-Context (AaC) tasks, we provide a set of prompt templates that annotators follow when writing questions. These templates cover three families of AaC tasks: (i) *Grounded QA*, which focuses on localizing and describing the sound source in the image (e.g., counting possible sources, attributes, and spatial relations); (ii) *Compositional Audio*, which includes sequential and multi-source audio and requires reasoning about event order or different sound components; and (iii) *Discourse*, which pairs long-form speech with structured visual content (e.g., diagrams or slides) and asks for concept identification or correspondence. Each template is further labeled with a coarse difficulty level (*easy*, *medium*, or *hard*), indicating the expected reasoning depth and guiding annotators toward balanced coverage across complexity levels. Figure 8 illustrates representative templates used during annotation.

Categories of Cross-Modal Matching Tasks As described in Section 3.5, the Cross-Modal Matching task is mainly organized using the AudioSet [18] taxonomy. We follow its category–subcategory structure when designing the matching items: annotators construct questions within the same category and include distractors drawn from the same subcategory to ensure fine-grained discrimination. All categories and subcategories used in our benchmark are listed in Table 7.

10.3. Design of Multi-Turn Task Graphs of Interaction-Level Tasks

Figure 9 presents the multi-turn task graphs and interaction state trajectories for the three representative subtasks: navigation, music reproducing, and puzzle solving. Each task begins at an initial state defined by its task graph. At every turn, the model receives multimodal observations—such as images, speech queries, or audio clips—and produces a textual response. The response is evaluated by an LLM-based judge, which determines whether it satisfies the conditions for transitioning to the next node in the task graph.

Table 7. Categories, subcategories, and question templates used for the cross-modal matching task. Annotators instantiate the corresponding template when constructing each audio–image matching item.

Category	**Subcategory**	**Question Template**
Animal	dog, cat, bird, canidae, sheep	Which image shows the animal that is making this sound?
Music Emotion	happy, epic, peace, romantic, sad	Which image best matches the overall mood of this piece of music?
Human Voice	human_voice, human_group_actions, human_sounds_other	Which image best matches the human activity or vocalization in the audio?
Instrument	string_violin, string_viola, string_cello, string_double_bass, string_harp, woodwind_flute, woodwind_piccolo, woodwind_oboe, woodwind_english_horn, woodwind_clarinet, woodwind_bassoon, brass_french_horn, brass_trumpet, brass_trombone, brass_tuba, keyboard_piano, keyboard_organ, keyboard_harpsichord, percussion_timpani, percussion_xylophone, percussion_glockenspiel, percussion_marimba, default	Which image shows the musical instrument that is producing this sound?
Vehicle	vehicle_boat, vehicle_motor, vehicle_rail, vehicle_aircraft	Which image shows the type of vehicle that is producing this sound?
Object Material	ceramic, glass, iron, plastic, wood	Which image best matches the dominant material of the object producing this sound?

Evaluation Prompts for Open-ended Tasks

You are an AI assistant responsible for judging whether a model's answer correctly addresses a given audio-visual joint reasoning question.
For each evaluation instance, you will receive: 1. **Question** (text only) 2. **Model Response** 3. **Ground-Truth Answer**
Your task is to determine whether the model's response aligns in meaning with the ground truth.
A response should be marked **correct** if it conveys the same meaning as the ground truth, even if phrased differently.
A response should be marked **incorrect** if it contradicts the ground truth or omits key information or introduces incorrect or irrelevant details.
Your evaluation must output: – **Correctness**: "correct" or "incorrect" – **Reason**: A brief justification

JSON Output Instruction (Translated):
Please return only the following JSON (do not output any extra text):

```
{
  "verdict": "correct|incorrect",
  "reason": "A brief explanation within 50 characters"
}
```

– question: {question} – gt: {reference} – prediction: {prediction}

For Audio-Grounded QA

Example 1 Question: Is the creature in the picture making this sound with its eyes open or closed?

Answer: The creature is making the sound with its eyes open.

Groundtruth: Open

Judge Result: Correct

Example 2 Question: How many animals in the picture can make this sound?

Answer: There are two animals that can make this sound.

Groundtruth: Two

Judge Result: Correct

Example 3 Question: What is the color of the hat worn by the person closest to the operator making this sound?

Answer: The person is wearing a blue hat.

Groundtruth: Yellow

Judge Result: Incorrect

Figure 4. **LLM judging prompt used for open-ended and interaction-level evaluation.** The judge receives the question, the model's answer, and the reference answer (or state-transition success criterion for interaction tasks), and determines whether the model's output should be marked as *correct*.

Original Questions	Spoken Form Questions
Segment BD and AE intersect at C, as shown, $AB=BC=CD=CE$, and $\angle A=\frac{5}{2} \angle B$. What is the degree measure of $\angle D$?	Segment BD and segment AE intersect at point C, as shown. AB equals BC equals CD equals CE, and angle A equals five halves times angle B. What is the degree measure of angle D?
What is <image 1> known as?	What is the information in the figure known as?
Societies practise the concept of <image 1> to maintain _____.	Societies practise the concept of the information shown in the figure to maintain what?

Figure 5. Examples of written questions and their converted spoken forms used in our *Audio-as-Instruction* tasks. To obtain natural, TTS-friendly prompts, we use GPT-4o to rewrite the questions into spoken-form utterances while preserving their original semantics.

Prompt Used for Quality-Based Downsampling

Please rate the quality of the following multimodal QA on a scale of 0–10, considering its overall content and context. Use the following criteria:

1. **Accuracy**: Are the facts, arguments, or explanations correct and credible? Identify any errors or misinformation.
2. **Logical Coherence**: Does the dialogue follow a logical flow? Are the points connected clearly?
3. **Clarity**: Is the language easy to understand and effectively expressed?
4. **Relevance**: Does each part of the dialogue meaningfully contribute to the topic?
5. **Engagement and Depth**: Is the discussion sufficiently deep, analytical, or insightful given the context?

Provide an overall score (0–10) with concise feedback on each aspect and the reasoning behind your rating.

Scoring Guidelines:

- **0–3**: Major issues (inaccuracies, poor coherence, unclear language, off-topic).
- **4–6**: Some strengths but contains errors, unclear reasoning, or lacks depth/relevance.
- **7–8**: Good quality with mostly accurate, coherent, and relevant content but minor issues.
- **9–10**: High-quality dialogue—accurate, logically sound, clear, relevant, and engaging.

The multimodal QA is given as:

```
{
    'Image': The image given,
    'Question': <question>,
    'Answer': <answer>
}
```

Please put your score within `\boxed{}`.

Figure 6. Prompt used during Quality-Based Downsampling in the data construction pipeline. We apply a quality-based downsampling procedure to ensure balanced and high-quality coverage across the AaI sub-tasks. Each candidate sample is evaluated by GPT-4o using a unified scoring prompt, which rates accuracy, logical coherence, clarity, relevance, and engagement/depth on a 0–10 scale.

Prompt Template for Spoken-Form Conversion

Instruction: You will be given a question. Convert it into a natural spoken form, making LaTeX formulas and mathematical symbols understandable when read aloud. If the question does not contain any LaTeX or special symbols, return the original text as-is. Do not provide any additional explanation or commentary.

Conversion Rules:

- Preserve all non-mathematical text exactly as written.
- Convert LaTeX formulas and mathematical symbols into clear spoken English.
- Identify the type of question:
  - For multiple-choice or fill-in-the-blank questions, rewrite them as complete, naturally phrased questions suitable for spoken delivery. Remove any brackets or underlines.
  - For incomplete statements (e.g., "The sum is equal to"), reformulate them into full questions (e.g., "What is the sum?"), while preserving their mathematical intent.
- If the question contains no LaTeX or symbols, return it verbatim.
- For placeholders such as `<image 1>`:
  - If it functions grammatically (e.g., as a subject or object), replace it with descriptive phrases such as "the data shown in the figure" or "the values in the image."
  - If it is a secondary or parenthetical reference, remove it entirely.
- Return only the converted spoken-form question. Do not include formatting, symbols, or meta-commentary.

Example Input:
Unreadable Question

Expected Output:
Expected Spoken Form Question

Figure 7. Prompt used to convert written questions into natural spoken form for TTS-based multimodal evaluation.

Prompt Template for Audio-as-Context Tasks

Grounded QA

1. (Attribute, easy) Is the animal making this sound with its mouth open or closed?
2. (Attribute, easy) What color is the object that can make this sound?
3. (Count, easy/medium) How many objects in the picture can make this sound?
4. (Spatial Relation, medium) What object is above/below the object making this sound?
5. (Spatial Relation, medium) What is on the road on the other side of the sound source?

Compositional audio

1. (Sequential audio, medium) Based on the images and sounds, what actions did the person in the pictures perform? (with options)
2. (Sequential audio, medium/hard) Based on the image and sound, how many times did the person in the picture hit the bowl against the sink?
3. (Sequential audio, hard) Based on the image and the audio clip, what can you infer about the reason for the screaming sound?
4. (Multi-sourced audio, medium) Which musical instrument shown in the image is not present in the audio?
5. (Multi-source audio, hard) Does the person playing the main melody in the audio wear glasses?

Discourse

1. (Audio-conditioned visual query, medium) The speaker used a word to describe the process indicated by the red arrow in the figure. What is this word?
2. (Image-conditioned audio query, medium) The speaker highlighted one modification as a good marker of aging cells. Which structure in the figure corresponds to this?
3. (Integrated reasoning, hard) Based on the image and the audio, what is the meaning of the computer icon in this context?

Figure 8. **Prompt templates for Audio-as-Context tasks.** The templates are organized into three families—Grounded QA, Compositional Audio, and Discourse—and each is annotated with a coarse difficulty level (*easy*, *medium*, or *hard*) to guide balanced question construction.

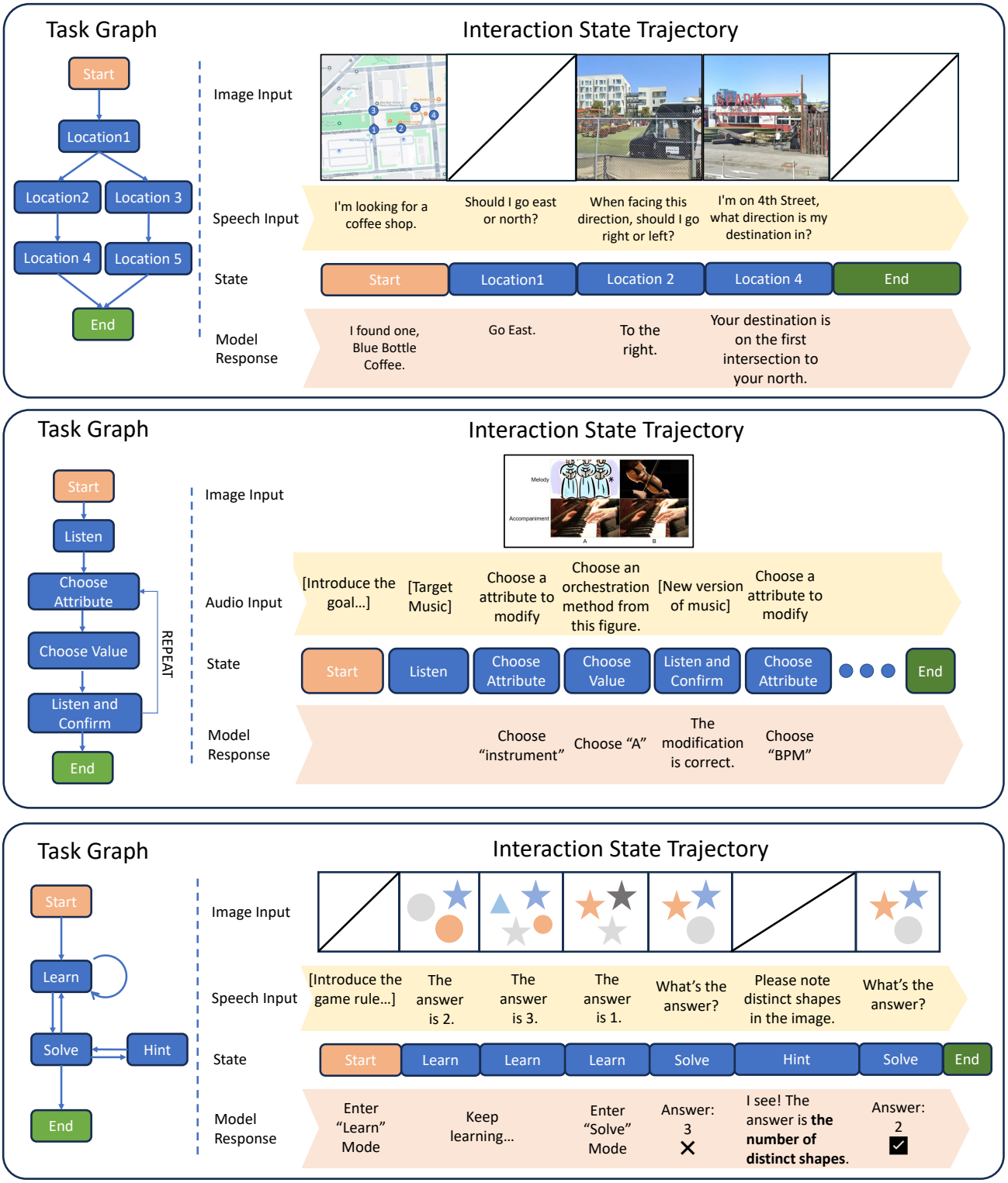

Figure 9. Multi-turn task graphs and interaction trajectories for the three interaction-level tasks.