PEAP: Proactive Embodied Action Sequence Planning with Joint Understanding of Vision and Audio Perception

Tianwei Lan[1,*], **Jiaqi Wu**[1,*], **Zeming Liu**[2,*], **Zhaoxin Fan**[2], **Haifeng Wang**[3], **Yuhang Guo**[1†]
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]School of Computer Science and Engineering, Beihang University [3]Baidu Inc.
{twlan,jiaqiwu,guoyuhang}@bit.edu.cn
{zmliu,zhaoxinf}@buaa.edu.cn wanghaifeng@baidu.com

Abstract

Embodied Action Sequence Planning focuses on the capability of embodied agents to implement action planning via environmental perception. This technology enables diverse intelligent assistance for real-world scenarios such as home and office environments. To address the limitations of existing embodied agents in meeting the requirement for proactivity and achieving joint understanding of visual and audio information, this study investigates the ability of embodied agents to proactively provide assistance through action sequence planning based on joint understanding of vision and audio perception without explicit human instructions. Correspondingly, we propose **PEAP**, the first multimodal proactive embodied action sequence planning dataset. We evaluate the performance of multiple Large Language Models on the PEAP dataset. The results demonstrate that these models still exhibit significant deficiencies on this task particularly lacking accurate environmental perception capabilities. Furthermore, ablation experiment and replacement experiment further corroborate that the joint understanding of multimodal information can significantly improve the models' performance on proactive embodied action sequence planning task. Our dataset and code are publicly available[1].

1 Introduction

Embodied action sequence planning technology focuses on embodied agents' environmental perception capabilities, aiming to achieve natural interactions between the agent, the environment, and humans, while providing diverse assistance for practical settings such as home and office (Yang et al., 2025a; Duan et al., 2022; Wu et al., 2023). This technology can offer auxiliary support to users

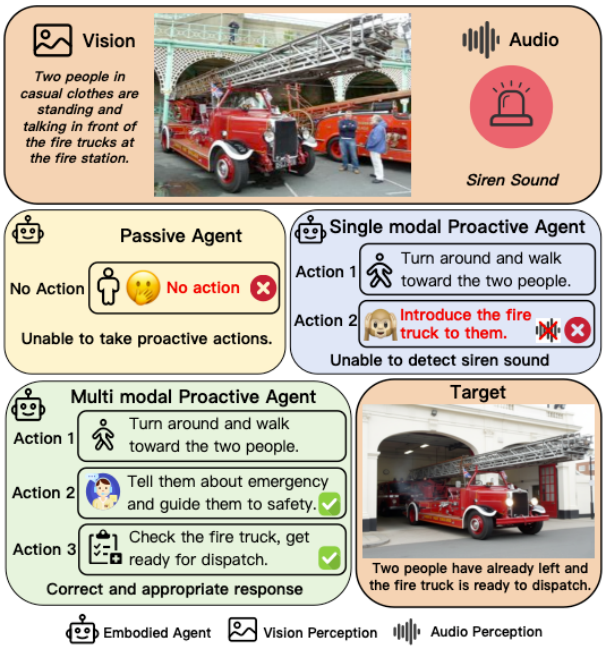

Figure 1: PEAP: Proactive Embodied Action Sequence Planning with Joint Understanding of Vision and Audio Perception

in work scenarios. For instance, answering task operation questions, retrieving task-related information, and responding to users' basic needs during task progression (Fan et al., 2025; Ren et al., 2024). However, in scenarios requiring the embodied agent to make proactive response decisions based on environmental conditions, like emergency handling and prediction of potential needs, passive embodied action planning struggles to meet the requirements because it relies on explicit instructions to trigger responses (Zhao et al., 2025a; Yang et al., 2025b). Therefore, we need embodied agents to proactively perceive environmental states, predict users' potential demands, and initiate proactive action planning to achieve efficient responses.

To address the aforementioned limitations of passive embodied action planning, the proactive embodied action planning mechanism should achieve

[*]Equal contribution
[†]Corresponding author
[1]https://github.com/BITHLP/PEAP

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23118–23138

efficient responses to such needs (Kraus et al., 2020; Gao et al., 2022; Zhang et al., 2025). Compared with passive embodied agents, proactive embodied action planning can autonomously assess the surrounding environment and occurring events, thereby determining whether action is required and the specific course of action, eliminating the intermediate link of human-initiated instructions. Consequently, this technology reduces users' operational burden, identifies users' unexpressed potential needs, and improves service execution efficiency and user interaction experience (Zhao et al., 2025b; Lu et al., 2024).

Proactive embodied action planning agents need to possess the capability of multimodal perception and joint understanding. When conducting scene comprehension, they must synchronously integrate multimodal inputs such as vision, audio, and text, fully capture visual details, sound signal features, and linguistic interaction information in the scene, and achieve comprehensive perception and judgment of complex scenarios. On the other hand, Proactive embodied action planning also needs the ability of in-depth mining of audio information to perceive scene attributes and sound events contained in background audio, and leverage such environmental acoustic information for scene understanding and user demand prediction, rather than focusing solely on the semantic content of the speaker.

To address the aforementioned requirements, this paper proposes PEAP, a dataset for proactive embodied action planning. The dataset includes real-life scenarios and corresponding audio data in each scenario. In each data sample, visual scene information and audio information collaboratively construct the context for the agent. The agent is required to output proactive assistance schemes or response content that aligns with scene needs through the joint understanding of vision and audio information. As illustrated in Figure 1, agents with multimodal joint perception capabilities can respond correctly in emergency scenarios, while passive agents and single-modal agents fail due to their inability to take proactive actions and detect sound information.

To verify the effectiveness and practicality of PEAP, we test multiple large language models on this dataset, categorizing all models into cascaded schemes and end-to-end schemes. By annotating and constructing a training dataset for the Evaluation Model (EM), we complete the fine-tuning and optimization of the EM and establish a quantitative evaluation system with both objectivity and stability. Through the evaluation experiments, we reveal the capability characteristics and performance boundaries of these models in proactive embodied action planning tasks. Additionally, ablation and replacement experiments confirm the necessity of multimodal information joint understanding for proactive embodied action sequence planning tasks.

The main contributions of this paper can be summarized as follows:

(1) Innovative Task: To our knowledge, we are the first to investigate the capability of embodied agents to deliver proactive assistance through action planning in multimodal scenarios.

(2) New Dataset: We construct the first multimodal proactive embodied action planning dataset **PEAP**, which surpasses existing datasets in terms of proactivity, multimodality, information joint understanding, and the number of scenarios.

(3) Systematic and Comprehensive Evaluation: We conduct comprehensive assessment of LLMs and in-depth analysis of the actions they take, revealing the deficiencies of LLMs in environmental perception as well as the importance of joint understanding of vision and audio for this task.

2 Related Work

2.1 Action Planning Agents

With the continuous advancement of large language models (LLMs) (DeepSeek-AI et al., 2024; OpenAI et al., 2023; Grattafiori et al., 2024; Yang et al., 2025) and embodied action planning technology (Tang et al., 2025; Nuzzi et al., 2024; Fan et al., 2025; Ren et al., 2024), action planning agents have been practically applied in multiple domains of daily life and work. Existing studies have explored the application scenarios of action planning agents, covering areas such as home assistants(Ahn et al., 2022), computer operation (Qian et al., 2023; Qin et al., 2023), and personal assistants(Yang et al., 2024). The research focus of these studies is centered on enhancing the performance of LLM-based Agents in core capability dimensions, specifically including task planning(Yao et al., 2022), logical reasoning(Mialon et al., 2023), and multi-agent collaboration capabilities(Zhang et al., 2023; Divekar et al., 2019).

Although the aforementioned studies have constructed a variety of task-specific frameworks and

DataSet	Proactive	Multimodal			Audio & Vision Joint Understanding	Scenarios
		Text	Audio	Vision		
Mem-PAL(Huang et al., 2025)	✘	✔	✘	✘	✘	100
PACHAT(Fu et al., 2025)	✘	✘	✔	✘	✘	21
OphGLM(Deng et al., 2024)	✘	✔	✘	✔	✘	1
LlamaPIE(Chen et al., 2025)	✔	✘	✔	✘	✘	5
ProactiveBench(Lu et al., 2024)	✔	✔	✘	✘	✘	3
PROASSIST(Zhang et al., 2025)	✔	✔	✘	✔	✘	4
DialFRED(Gao et al., 2022)	✔	✔	✘	✔	✘	25
PEAP	✔	✔	✔	✔	✔	**122**

Table 1: Comparison between PEAP and other datasets

datasets, they all focus on passively responsive action planning agents. Such Agents rely on explicit user instructions to determine the timing of task initiation and lack the core capability to perceive scenario information and proactively provide assistance autonomously (Zhao et al., 2025a; Yang et al., 2025b). In contrast, this study focuses on investigating the assistance capability of LLM-based action planning agents in proactive action planning scenarios, filling the gap in the "initiative" dimension existing in these works.

2.2 Proactive Embodied Agents

Proactive embodied agents can act without waiting for explicit instructions from humans and possess autonomous initiative. Some existing studies focus on endowing agents with active interaction capabilities. For instance, enabling them to proactively ask users questions or initiate new topics (Gao et al., 2022; Zhang et al., 2025). This allows for clarifying originally ambiguous semantics during interaction or putting forward innovative suggestions. Other studies focus on active operation assistants, covering both computer-based and mobile operation assistants (Zhao et al., 2025b; Lu et al., 2024). By monitoring screen information of electronic devices, such assistants perceive users' intentions and real-time events, and then proactively provide assistance. In addition, some studies have put forward targeted insights regarding the construction of evaluation systems for Proactive Embodied Agents, providing references for the improvement of evaluation standards in this field (Liu et al., 2025).

However, existing research related to Proactive Embodied Agents has not focused on full-modal scenarios with simultaneous input of audio, visual, and text information (Hassan et al., 2024; Zhao et al., 2025a; Fan et al., 2024; Kim et al., 2025), whereas real-world scenarios often contain information from these three modalities. Meanwhile, some works merely use audio as a tool for inputting and outputting semantic information in task settings, ignoring the vast amount of information in the audio background that can indicate sound events or the main subjects of the scene (Jolibois et al., 2023; Shubham et al., 2022; Yang et al., 2025a). In practical situations, this may lead to incorrect judgments about the scene and thus the inability to provide appropriate responses. To address this deficiency, our work focuses on researching the ability of Proactive Embodied Agents to jointly understand multimodal information and the impact of background audio information on the responses made by Agents.

We have listed a comparison between PEAP and other datasets in terms of Proactive and Multimodality in the Table 1. The proposed PEAP is the only dataset that achieves full-modal input and proactive action planning.

3 Data Construction

We introduce the Task Definition and the construction process of the dataset, including Data Collection, Data Filtering & Annotation and Quality Control.

3.1 Task Definition

Our task aims to evaluate embodied agents' capability to provide proactive assistance via action planning in multimodal scenarios. Embodied agents receive three types of scene information: visual scene information V, audio scene information A, and textual task description T, where V and A have specific logical correlations. Agents are required to identify scene demands and plan actions through multimodal joint perception of V and A.

Drawing on prior works in embodied intelligence (Ma et al., 2024; Shridhar et al., 2019) and robotic manipulation (Tellex et al., 2011; Bai et al., 2025), we define three action categories for embodied agents as shown in Table 2.

Each action output follows the format: [MajorCategory][Subcategory]SpecificContent. For instance, [Manipulation][Grab] Grab a towel from the shelf in the bathroom.

The task is formally defined as:

$$ActionSequence(Action_1, Action_2, ..., Action_n) = EmbodiedAgent(V, A, T) \quad (1)$$

where $Action_i$ denotes the i-th action in the sequence, V = Vision, A = Audio, T = Task description.

3.2 Data Collection

To address the deficiency in existing embodied agent action planning datasets that lack joint understanding of Audio and Vision information, we extract data of images and audio from existing public datasets. The adopted image datasets include SUN397(Xiao et al., 2010), MIT67 (Quattoni and Torralba, 2009) and Place365 (Zhou et al., 2017), while the audio datasets cover UrbanSound8K (Salamon et al., 2014), FSD50K (Fonseca et al., 2021) and ESC-50 (Piczak, 2015). Subsequently, we manually paired scene categories with audio categories to ensure that the paired audio has a reasonable possibility of occurring in the corresponding scene to achieve logical relevance between scenes and audio. A typical example is matching the fire station scene with the siren sound. Through pairwise matching, we constructed the initial scene-audio data pairs. To further ensure the natural consistency between scenes and audio, we additionally collected sound-containing video data from Ego4D (Grauman et al., 2021) and VGGSound (Chen et al., 2020) and incorporated it into the dataset of this study. We present some examples of scene-audio pairings in Table 3.

3.3 Data Annotation & Quality Control

During the data filtering and annotation phase, we employed the image captioning model JoyCaption[2] and the audio captioning model AudioFlamingo3[3]. With the ground truth labels of images and audio provided, the two models were driven to generate textual descriptions corresponding to the input images and audio, respectively. Subsequently, we adopted the DeepSeek-V3.2[4] model to perform data filtering based on the generated textual descriptions of images and audio, ensuring that the assistance task has a clear execution direction under the condition of specific scenes combined with concrete sound events. For the filtered data that meet the directional requirements, we further leveraged DeepSeek-V3.2 to annotate the core content of the assistance task.

To ensure data quality, we adopt a quality control strategy involving manual review, where experts audit the data annotated by large models to guarantee the relevance between scenes and audio as well as annotation accuracy. During the manual verification of data annotated by large models, human annotators can not only choose to accept or reject the annotations but also revise the original annotation content generated by the model, so as to avoid anchoring bias caused by model annotation. Ultimately, 10% of the final obtained data is randomly sampled for manual inspection, with an annotation consistency rate of over 95% (Lu et al., 2020). It should be clarified that the answers we annotated represent the general content of the appropriate assistance directions corresponding to each scene combined with audio, serving as a description of the assistance directions rather than specific action sequences. Therefore, action sequences generated by multiple different models can all meet the requirements and be accepted after being judged by the evaluation model. This ensures that our task can, to a certain extent, adapt to situations where different reactions occur in the real world for the same scenario, as long as consistency with the direction of the answers is maintained, as shown in Figure 9. In addition, our data includes a number of no-action-needed cases, for which the annotated answer is to take no assisting action. During testing, the model should choose not to perform any assisting action for these cases, which will be regarded as correct. Through the above annotation process, we can see that PEAP is a controlled, normative benchmark rather than a human-grounded simulation of real-world proactive behavior, and it accommodates diverse possibilities in the real world.

The prompt formats for textual description gen-

[2] https://hf-mirror.com/fancyfeast/llama-joycaption-beta-one-hf-llava

[3] https://hf-mirror.com/nvidia/audio-flamingo-3

[4] https://api-docs.deepseek.com/

Action Category	Description	Examples
Movement	Positional movement	Walking forward, Turning around
Manipulation	Object interaction	Grabbing a towel, Opening a door
Conversation	Verbal behavior	Inquiring about something, Raising a topic

Table 2: Three action categories for embodied agents and their description and examples.

Scenes	Audios
Airport	A baby is crying
	A person coughs
	An aircraft can be heard
Classroom	A door is slammed
	A female teacher or student is speaking
	Footsteps are heard
Gym	A person is running on a treadmill
	A person is shouting during a workout
	Footsteps are heard on the gym floor

Table 3: Examples of several scenarios and sound events that may occur in them.

Model	SR	SER	AP	Overall
Qwen3-8B	29.3	62.0	52.7	61.3
Llama3-8B	64.0	50.7	65.3	64.7
Deepseek-api	82.0	77.3	68.7	92.7
Qwen3-8B-FT	74.0	80.0	80.7	89.3
Llama3-8B-FT	93.3	92.0	82.0	88.7
OurMethod	**98.0**	**99.3**	**96.7**	**98.7**

Table 4: Evaluation models' performance on the evaluation test dataset.

eration, data filtering, and answer annotation are detailed in Appendix A.1. Finally, we filtered and annotated the PEAP dataset, which covers 122 scenes and contains a total of 19963 data entries. The format of the data samples and the details of the dataset are presented in the Appendix B.

3.4 Quality Validation

Following previous work (Lu et al., 2020), we randomly sampled 2000 data entries from it and manually evaluated whether each sample has a clear assistance direction and whether the annotated assistance content is accurate. The evaluation results show that the valid acceptance rate of the dataset reaches as high as 98.9%.

4 Evaluation

4.1 Metrics

Following previous work (Xie et al., 2020; Bilen et al., 2020; Lu et al., 2024), we set four distinct metrics to evaluate the quality of the action sequences generated by the model. (1) Scene Recognition (SR): whether the model can identify the current scene through visual information. (2) Sound Events Recognition (SER): whether the model can recognize the ongoing event through audio input. (3) Assistance Provide (AP): whether the model provides assistance to humans in the scene. (4) Overall: This metric measures the consistency between the action sequences generated by models and the reference answers of each data sample, and directly reflects whether models complete the PEAP task.

For the three aforementioned sub-metrics, we adopt the evaluation model introduced in the next subsection to conduct independent judgments. For any given model response, the PEAP metric is scored as 1 only if the judgment results of all three sub-metrics meet the requirements; otherwise, it is scored as 0.

4.2 Evaluation Model

We introduce evaluation models to score the responses generated by the model. For the four metrics mentioned earlier, the evaluation model assigns a binary score of 0 or 1 to determine whether the response meets the answer requirements for a specific metric. To improve the scoring accuracy of the evaluation model, we manually annotated training data for further training of the evaluation model. Meanwhile, we incorporate a human inspection step into the evaluation scheme to ensure accuracy further.

We select data samples from all scenes in PEAP and use these data to test all of our test models and obtain the output. We recruited master's and doctoral students specializing in artificial intelligence to conduct manual annotation and scoring of the dataset based on four evaluation metrics, eventually forming approximately 6k evaluation data samples with about 1.5k corresponding to each metric.

We test the scoring accuracy of Qwen3-8B, Llama3-8B, and DeepSeek as evaluation models. To improve the judgment accuracy of Qwen3-8B and Llama3-8B, we adopt the LoRA method to

fine-tune the two models on the training set of the evaluation data. The test results of all evaluation models are shown in Table 4. It can be seen that the judgment accuracy of Qwen3-8B and Llama3-8B has been significantly improved after fine-tuning.

4.3 Evaluation Method

In the actual evaluation process, we use DeepSeek, Qwen3-8B-FT, and Llama3-8B-FT simultaneously to score the model outputs. If the three models yield the same score for a single data sample, we directly accept the result. Otherwise, we manually score the data sample to ensure the accuracy of the scoring results. The results in Table 4 show that our method achieves high accuracy.

5 Experiment

5.1 Models

All the test models adopted in this study are divided into two schemes, namely the cascade scheme and the end-to-end scheme. In the cascade scheme, we first leveraged the aforementioned JoyCaption and AudioFlamingo3 models to generate textual descriptions for image and audio data, respectively, and then fed the scene information into text-modal models. It should be noted that during the textual description generation phase, no ground-truth labels of the data were provided to the captioning models. This setting will lead to recognition deviations in the models under certain circumstances, thereby simulating the error propagation problem inherent in cascade models. The models tested using the cascade scheme include Llama3-8B (Grattafiori et al., 2024) and the 0.6B, 8B, and 32B versions of Qwen3 (Yang et al., 2025). The format of the test prompts for cascade models is detailed in the Appendix A.2.

In the end-to-end scheme, we directly input image and audio information into multi-modal models, relying on the models' inherent cross-modal understanding capabilities to parse image and audio content, while describing task information in textual form. The models tested using the end-to-end scheme include Mini-omni(Xie and Wu, 2024), MiniCPM(Hu et al., 2024), Stream-omni(Zhang et al., 2025), Qwen3-omni(Xu et al., 2025), VITA(Fu et al., 2024), GPT-4o(Hurst et al., 2024), o4-mini[5], Gemini-3 pro[6]. The format of the test prompts for end-to-end models is detailed in the Appendix A.3.

5.2 Main Results

Table 5 presents the performance of various models on the PEAP task. Overall, there exists a significant performance gap among models: top-performing models include Qwen3-32B, Qwen3-omni and Gemini-3 Pro, all achieving overall scores exceeding 70, while underperforming models such as Qwen3-0.6B and Mini-omni yield notably low overall scores.

Analysis of the three sub-metrics reveals that the final model performance is constrained by the weakest sub-metric. This is because accurate action planning relies on the synergistic integration of Scene Recognition, Sound Events Recognition and effective Assistance Provide. A low score in any sub-metric hinders the generation of correct action plans. Take Qwen3-0.6B as an illustration. This model performs moderately in Scene Recognition but exhibits an extremely low Assistance Provide score, ultimately leading to poor overall performance on the PEAP task. This pattern is most prominent in the VITA model, despite its high accuracy in Scene Recognition under multimodal input, its deficiencies in Sound Events Recognition and Assistance Provide result in task failure. In contrast, top-performing models Qwen3-omni and Gemini-3 Pro demonstrate robust performance across all three sub-metrics, which contributes to their high overall scores on the PEAP task. Case studies on specific data are provided in Appendix B.3.

6 Further Analysis

We conduct in-depth analysis to address three core research questions (RQs) on proactive embodied action planning. These questions explore the rationality of model-generated action sequences, key factors influencing model performance, and the necessity of multimodal joint understanding.

6.1 RQ1: Will models take redundant actions to improve the task completion rate?

Through this experiment, we hope to identify whether the actions output by the models can hit the core points specified in the reference answers (Driess et al., 2023), that is, whether the models accurately understand the key demands for assistance in the scene. On the other hand, we also evaluate

[5] https://platform.openai.com/docs/models/o4-mini

[6] https://deepmind.google/models/gemini/pro/

Model		PEAP Task				Action Analysis			
		SR	SER	AP	Overall	Precision	Recall	Macro-F1	Micro-F1
Cascade	Llama3-8B	80.0	58.5	45.3	41.0	49.6	44.2	40.3	46.7
	Qwen3-0.6B	30.9	16.7	7.5	6.5	12.0	14.5	9.5	13.1
	Qwen3-8B	81.4	77.9	73.5	68.8	53.9	55.7	49.2	54.8
	Qwen3-32B	84.5	79.9	76.6	71.7	50.7	60.0	50.5	54.9
End2End	Mini-omni	14.6	8.5	6.9	5.9	8.2	3.1	2.0	4.5
	MiniCPM	**89.5**	38.9	42.0	35.8	37.8	37.3	33.0	37.5
	Stream-omni	60.8	28.4	23.9	21.6	27.4	32.6	23.4	29.8
	Qwen3-omni	88.8	80.9	79.7	76.0	79.4	82.4	75.2	80.8
	VITA	83.7	19.9	16.5	14.4	32.9	42.2	31.0	36.9
	GPT-4o	67.3	56.4	61.4	56.4	45.1	54.0	45.8	49.2
	o4-mini	72.3	61.4	67.3	61.4	54.3	60.2	55.8	57.1
	Gemini-3 Pro	87.6	**84.7**	**89.2**	**80.6**	**85.5**	**90.1**	**80.9**	**87.8**

Table 5: Performance Comparison of Cascade and End2End Models on PEAP Task, Modality Ablation and Modality Replacement Experiments. SR stands for "Scene Recognition". SER stands for "Sound Events Recognition". AP stands for "Assistance Provide". Precision represents the proportion of actions predicted by the model that match the actions in the answer. Recall represents the proportion of actions in the answer that are correctly predicted by the model. F1 is the harmonic mean of Precision and Recall.

whether the models generate a large number of redundant actions irrelevant to the final goals (Zhu et al., 2021), and whether they achieve relatively high scores through extensive guessing (Yao et al., 2022). As shown in Figure 8 in the appendix, the model outputs redundant actions irrelevant to the scene semantics in this example.

The experiment of Action Analysis in Table 5 conducts a more detailed analysis of the actions output by the models. We manually split the reference answers of the test data into individual actions, and also split the action sequences output by the models into corresponding single actions. On this basis, we separately calculate the proportion of actions in the reference answers that are successfully output by the models, and the proportion of actions output by the models that match the reference answers. These two metrics are represented by Recall and Precision, respectively, and the F1 score is calculated accordingly.

The experimental results show that Gemini-3 Pro and Qwen3-omni achieve relatively high scores, while the scores of Qwen3-32B and o4-mini are not as high as those in the PEAP task experiment. This indicates that these two models may omit some steps that do not affect the final result judgment during the action prediction process. In addition, the Recall of Qwen3-32B is significantly higher than its Precision, which suggests that the model outputs a large number of actions during prediction. Although this enables the model to hit more actions in the ground truth, it also reduces the Precision score. This means that it improved its score on the PEAP task by outputting some redundant actions.

6.2 RQ2: Does multimodal information input help models improve their performance?

This experiment aims to verify that the joint understanding of multi-modal information effectively improves the action planning performance of models. To this end, we design a modality ablation experiment (McKinzie et al., 2024), remove either the audio information or the vision information from the original test set, respectively, and observe the changes in the action planning capabilities of all models under the conditions of Only Vision and Only Audio, including whether the capabilities decline and to what extent the decline occurs (Ao et al., 2025). If the action planning scores of models show a significant decrease under the single-modal input condition, it indicates that single-modal information cannot provide sufficient information for models in multi-modal scenarios to determine which actions to take to assist humans (Garrett et al., 2020).

As shown in Table 6, all models exhibit significant degradation in action planning performance under single-modal input, with substantial score declines regardless of baseline capabilities.Under Only Vision, models lack audio semantic cues, failing to effectively identify event properties and relying solely on visual features for reasoning. Un-

Model	V + A	Only V	Only A
Llama3-8B	41.0	30.0	1.1
Qwen3-0.6B	6.5	5.1	1.9
Qwen3-8B	68.8	36.7	1.7
Qwen3-32B	71.7	37.6	2.6
Mini-omni	5.9	2.1	0.9
MiniCPM	35.8	17.7	2.3
Stream-omni	21.6	11.6	2.1
Qwen3-omni	76.0	39.3	2.0
VITA	14.4	9.1	2.0
GPT-4o	56.4	21.7	3.0
o4-mini	61.4	38.5	2.1
Gemini-3 Pro	80.6	40.0	12.6

Table 6: Performance of Models Under Single Modality Inputs. "Only V" denotes inputting only visual information, and "Only A" denotes inputting only audio information.

Model	No Repl.	Repl. V	Repl. A
Llama3-8B	41.0	8.8	14.3
Qwen3-0.6B	6.5	6.2	5.9
Qwen3-8B	68.8	9.7	29.2
Qwen3-32B	71.7	9.9	30.3
Mini-omni	5.9	5.2	3.1
MiniCPM	35.8	6.3	10.8
Stream-omni	21.6	4.3	18.9
Qwen3-omni	76.0	7.7	27.7
VITA	14.4	6.0	10.0
GPT-4o	55.4	8.7	19.8
o4-mini	61.4	9.1	20.1
Gemini-3 Pro	80.6	8.0	26.7

Table 7: Performance of Models Under Modality Replacement. "Repl. V" denotes replacing the visual information, and "Repl. A" denotes replacing the audio information.

der Only Audio, models cannot access object spatial distribution, severely limiting manipulation-oriented action planning. Both scenarios confirm that joint understanding of visual and audio information is a core enabler for multimodal action planning.Model performance under Only Vision is generally superior to that under Only Audio. Visual information encapsulates scene and object features, supporting partial basic planning but remaining far below the multimodal baseline; Only Audio leaves models nearly unable to generate valid action sequences. This demonstrates single-modal information cannot meet the PEAP task's cognitive demands, and multimodal joint understanding is a necessary prerequisite for proactive embodied action planning.

6.3 RQ3: Do models really achieve joint understanding of multimodal information?

This experiment is designed to verify models' genuine joint understanding capability of visual-audio multimodal information, ruling out the possibility of action plan generation via random guessing or single-modal dependence (Chao et al., 2025). Theoretically, the semantic connotation of a multimodal scenario is co-constructed by visual and audio information. A model with reliable joint understanding capability should perceive and integrate the semantic correlation between these two types of information (Su et al., 2024). Conversely, if a model only relies on single-modal information or generates outputs through random strategies, its performance will not fluctuate significantly with the alteration of core modal information (Zhong et al., 2024).

The experimental design is as follows: for each sample in the original test set, visual information replacement and audio information replacement are performed separately. A pure black image is used when replacing visual information, and pure noise audio is used when replacing audio information. Based on this, two modal mismatched test sets are constructed, namely the vision-replaced test set and the audio-replaced test set. By comparing the performance of the original models on these two mismatched test sets with the baseline performance on the original multimodal test set, the extent of performance degradation is analyzed to determine whether the models truly possess multimodal joint understanding capability.

Experimental results in Table 7 demonstrate that after replacing either visual or audio information, the scores of all models generally decrease substantially. This indicates that the process of models' joint understanding of multimodal information is disrupted by modality replacement. When reasoning with the replaced scene or sound information, the models generate outputs inconsistent with the original answers. Additionally, some models exhibit relatively minor score changes, such as Qwen3-0.6B. This suggests that such models may adopt a fixed output mode and fail to fully consider the impact of different sound events on scene semantics.

Model	Origin T	Simple T
Llama3-8B	41.0	33.0
Qwen3-0.6B	6.5	7.1
Qwen3-8B	68.8	59.7
Qwen3-32B	71.7	67.7
Mini-omni	5.9	4.3
MiniCPM	35.8	30.4
Stream-omni	21.6	15.3
Qwen3-omni	76.0	69.6
VITA	14.4	12.3
GPT-4o	55.4	40.1
o4-mini	61.4	50.8
Gemini-3 Pro	80.6	71.0

Table 8: Ablation experiment on descriptive text T. "Origin T" represents the result obtained using the original text description T, while "Simple T" represents the result obtained after removing the requirement for the agent to perform proactive action planning.

Model	SR	SER	AP	Overall
Llama3-8B	77.3	56.7	42.0	40.7
Qwen3-0.6B	30.0	12.0	8.7	7.3
Qwen3-8B	82.0	77.3	76.7	63.3
Qwen3-32B	84.7	80.0	77.3	71.3
Mini-omni	13.3	10.7	12.7	10.0
MiniCPM	96.7	55.3	61.3	54.7
Stream-omni	54.7	39.3	36.7	34.7
Qwen3-omni	94.0	89.3	79.3	75.3
VITA	86.0	38.0	36.7	36.7
GPT-4o	82.0	77.3	72.0	63.3
o4-mini	91.3	88.0	78.7	73.3
Gemini-3 Pro	90.0	88.7	83.3	82.0

Table 9: Human Evaluation on the performance of Cascade and End2End Models on PEAP Task.

6.4 RQ4: To what extent does task description T influence the performance of the model?

In the task setting of this study, text T is only used to define the role of the model, specifying the role that the model undertakes in the corresponding scenario, that is, to provide proactive assistance to users according to the actual situation of the scenario. It essentially defines the role attribute of the model rather than giving explicit instructions on the model's specific behaviors. This experiment aims to explore the degree of influence of such textual descriptions on model performance. To this end, we remove the content in the original textual description that requires the embodied agent to make autonomous decisions based on scenario information, and observe the changes in the model's performance metrics after removing this part of the description.

The experimental results are shown in Table 8. Most models show a slight performance decline after using Simple T. This indicates that providing a more accurate description of the tasks to be performed to embodied agents in the task description can better assist models in executing the Proactive Embodied Action Sequence Planning task, even without direct scenario-specific prompts for the models.

7 Human Evaluation

In this section, we present our human evaluation experiment. We uniformly select 150 data entries from the original test set, ensuring that each scenario is covered, and then conduct a human evaluation of the models' output results on this batch of data.

In Table 9, we conduct manual scoring on the models' action planning performance across four metrics, namely Scene Recognition (SR), Sound Events Recognition (SER), Assistance Provision (AP), and Overall, following the same scoring rules specified earlier. The results show high consistency with those in Table 5. Qwen3-omni, o4-mini, and Gemini-3 Pro remain the three best-performing models, while Qwen3-0.6B and Mini-omni still rank the lowest in terms of scores. This finding confirms the consistency of the two methods.

8 Conclusion

To address the limitations of existing embodied action planning research, namely the lack of proactivity and neglect of vision and audio joint understanding, we propose the PEAP task. This task constructs the environment by fusing visual and audio information, evaluating embodied agents' capability to provide proactive assistance based on vision and audio joint understanding. In-depth analysis of model-generated action sequences further verifies the models' ability to accurately identify human needs in scenarios without redundant actions. Modality ablation and replacement experiments confirm that multimodal joint understanding is crucial for models to provide assistance by action sequence planning.

Limitations

A consideration for future extension of this study relates to computational efficiency. In the current work, our focus has been on ensuring the quality of fine-grained joint understanding of multimodal information and the reliability of action planning reasoning. As such, targeted optimization for computational complexity and inference speed was not prioritized in the experimental setup. Specifically, when processing high-resolution visual inputs and long-duration audio sequences, the model may exhibit relatively higher inference latency, which suggests potential room for improvement in adapting to real-time deployment scenarios on edge devices with limited computing power. Moving forward, we plan to explore lightweight model architecture designs and efficient inference strategies, aiming to reduce computational overhead while preserving task performance, thereby further enhancing the practical deployment feasibility of the system.

Ethic Statement

Data Privacy Throughout the entire research process, we have strictly adhered to rigorous ethical standards, upholding the principles of transparency, fairness, and user privacy protection. The data utilized in the PEAP dataset is derived from publicly available datasets (including SUN397, MIT67, Place365, UrbanSound8K, FSD50K, ESC-50, Ego4D, and VGGSound) and has undergone meticulous processing to ensure the anonymization of any potential personal information. No identifiable private data of individuals is included in the dataset. All data collection, pairing, filtering, and annotation procedures are conducted in full compliance with relevant data protection and privacy regulations, minimizing any potential risks to user privacy.

Professional Annotation To guarantee the quality and accuracy of data annotation, we adopted a dual approach combining AI-assisted processing and manual validation. Initially, we leveraged mature models (JoyCaption, AudioFlamingo3, and DeepSeek-V3.2) for preliminary textual description generation, data filtering, and initial annotation. Subsequently, we engaged professional annotators with expertise in multimodal scene understanding to conduct manual review and refinement of the filtered data. These annotators received clear guidelines and fair compensation for their work, ensuring that the annotation process adheres to high standards of professionalism, objectivity, and responsibility while acknowledging their contributions appropriately.

Proactive Assistance Ethics In the design of the proactive embodied action planning task, we have fully considered the ethical implications of proactive assistance. The dataset scenarios and annotated target responses are constructed to prioritize user autonomy and safety, ensuring that the proactive actions suggested by the model do not violate user intentions, infringe on personal space, or pose safety risks. All assistance-oriented actions in the dataset are contextually appropriate, non-intrusive, and aligned with general social norms, promoting responsible and ethical application of proactive embodied AI technology.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments. This work issupported by the National Natural Science Foundation of China (Grant No.U21B2009).

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Shuang Ao, Flora D Salim, and Simon Khan. 2025. Emac+: Embodied multimodal agent for collaborative planning with vlm+ llm. *arXiv preprint arXiv:2505.19905*.

Shuanghao Bai, Wenxuan Song, Jiayi Chen, Yuheng Ji, Zhide Zhong, Jin Yang, Han Zhao, Wanqi Zhou, Wei Zhao, Zhe Li, and 1 others. 2025. Towards a unified understanding of robot manipulation: A comprehensive survey. *arXiv preprint arXiv:2510.10903*.

Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović. 2020. A framework for the robust evaluation of sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65. IEEE.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. 2025. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. VGGSound: A Large-scale Audio-Visual Dataset. *arXiv e-prints*, arXiv:2004.14368.

Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A Smith, and Shyamnath Gollakota. 2025. Llamapie: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13801–13824.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. DeepSeek-V3 Technical Report. *arXiv e-prints*, arXiv:2412.19437.

Zhuo Deng, Weihao Gao, Chucheng Chen, Zhiyuan Niu, Zheng Gong, Ruiheng Zhang, Zhenjie Cao, Fang Li, Zhaoyi Ma, Wenbin Wei, and 1 others. 2024. Ophglm: An ophthalmology large language-and-vision assistant. *Artificial Intelligence in Medicine*, 157:103001.

Rahul R Divekar, Xiangyang Mou, Lisha Chen, Maira Gatti De Bayser, Melina Alberio Guerra, and Hui Su. 2019. Embodied conversational ai agents in a multi-modal multi-agent competitive dialogue. In *IJCAI*, pages 6512–6514.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

Haolin Fan, Xuan Liu, Jerry Ying Hsi Fuh, Wen Feng Lu, and Bingbing Li. 2025. Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics. *Journal of Intelligent Manufacturing*, 36(2):1141–1157.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. 2024. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv e-prints*, arXiv:2408.05211.

Dongjie Fu, Xize Cheng, Linjun Li, Xiaoda Yang, Lujia Yang, and Tao Jin. 2025. Pachat: Persona-aware speech assistant for multi-party dialogue. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29313–29330.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2020. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *Proceedings of the international conference on automated planning and scheduling*, volume 30, pages 440–448.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2021. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *arXiv e-prints*, arXiv:2110.07058.

Sabit Hassan, Hye-Young Chung, Xiang Zhi Tan, and Malihe Alikhani. 2024. Coherence-driven multimodal safety dialogue with active learning for embodied agents. *arXiv preprint arXiv:2410.14141*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv e-prints*, arXiv:2404.06395.

Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Kehan Chen, Chuan Yu, Xubin Li, Tiezheng Ge, Wenxuan Wang, and Qin Jin. 2025. Mem-pal: Towards memory-based personalized dialogue assistants for long-term user-agent interaction. *arXiv preprint arXiv:2511.13410*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Simon Jolibois, Akinori Ito, and Takashi Nose. 2023. Multimodal expressive embodied conversational agent design. In *International Conference on Human-Computer Interaction*, pages 244–249. Springer.

Jiho Kim, Junseong Choi, Woosog Chay, Daeun Kyung, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2025. Propersim: Developing proactive and personalized ai assistants through user-assistant simulation. *arXiv preprint arXiv:2509.21730*.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 107–116.

Tianjian Liu, Fanqi Wan, Jiajian Guo, and Xiaojun Quan. 2025. Proactiveeval: A unified evaluation framework for proactive dialogue agents. *arXiv preprint arXiv:2508.20973*.

Jian Lu, Wei Li, Qingren Wang, and Yiwen Zhang. 2020. Research on data quality control of crowdsourcing annotation: A survey. In *2020 IEEE Intl Conf on dependable, autonomic and secure computing, Intl Conf on pervasive intelligence and computing, Intl Conf on cloud and big data computing, Intl Conf on cyber science and technology congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 201–208. IEEE.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, and 1 others. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv e-prints*, arXiv:2405.14093.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, and 1 others. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

Davide Nuzzi, Paul Cisek, and Giovanni Pezzulo. 2024. Planning-while-acting: addressing the continuous dynamics of planning and action in a virtually embodied task. *bioRxiv*, pages 2024–11.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 181 others. 2023. GPT-4 Technical Report. *arXiv e-prints*, arXiv:2303.08774.

Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6(3):1.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, and 1 others. 2023. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*.

Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1247–1254. IEEE.

Lei Ren, Jiabao Dong, Shuai Liu, Lin Zhang, and Lihui Wang. 2024. Embodied intelligence toward future smart manufacturing in the era of ai foundation model. *IEEE/ASME Transactions on Mechatronics*.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1041–1044.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *arXiv e-prints*, arXiv:1912.01734.

Kumar Shubham, Laxmi Narayen Nagarajan Venkatesan, Dinesh Babu Jayagopi, and Raj Tumuluri. 2022. Multimodal embodied conversational agents: A discussion of architectures, frameworks and modules for commercial applications. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 36–45. IEEE.

Ying Su, Zhan Ling, Haochen Shi, Cheng Jiayang, Yauwai Yim, and Yangqiu Song. 2024. Actplan-1k: Benchmarking the procedural planning ability of visual language models in household activities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14953–14965.

Weiliang Tang, Jia-Hui Pan, Wei Zhan, Jianshu Zhou, Huaxiu Yao, Yun-Hui Liu, Masayoshi Tomizuka, Mingyu Ding, and Chi-Wing Fu. 2025. Embodiment-agnostic action planning via object-part scene flow. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2086–2093. IEEE.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pages 1507–1514.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE.

Lin Xie, Feifei Lee, Li Liu, Koji Kotani, and Qiu Chen. 2020. Scene recognition: A comprehensive survey. *Pattern Recognition*, 102:107205.

Zhifei Xie and Changqiao Wu. 2024. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *arXiv e-prints*, arXiv:2408.16725.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-Omni Technical Report. *arXiv e-prints*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv e-prints*, arXiv:2505.09388.

Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. 2024. Viassist: Adapting multi-modal large language models for users with visual impairments. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pages 32–37. IEEE.

Fu-Chia Yang, Pedro Acevedo, Siqi Guo, Minsoo Choi, and Christos Mousas. 2025a. Embodied conversational agents in extended reality: A systematic review. *IEEE Access*.

Qinglong Yang, Haoming Li, Haotian Zhao, Xiaokai Yan, Jingtao Ding, Fengli Xu, and Yong Li. 2025b. Fingertip 20k: A benchmark for proactive and personalized mobile llm agents. *arXiv preprint arXiv:2507.21071*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. 2025. Stream-Omni: Simultaneous Multimodal Interactions with Large Language-Vision-Speech Model. *arXiv e-prints*, arXiv:2506.13642.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. 2025. Proactive assistant dialogue generation from streaming egocentric videos. *arXiv preprint arXiv:2506.05904*.

Yuheng Zhao, Xueli Shu, Liwen Fan, Lin Gao, Yu Zhang, and Siming Chen. 2025a. Proactiveva: Proactive visual analytics with llm-based ui agent. *arXiv preprint arXiv:2507.18165*.

Yuyang Zhao, Wentao Shi, Fuli Feng, and Xiangnan He. 2025b. Appagent-pro: A proactive gui agent system for multidomain information integration and user assistance. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6767–6771.

Linqing Zhong, Chen Gao, Zihan Ding, Yue Liao, Huimin Ma, Shifeng Zhang, Xu Zhou, and Si Liu. 2024. Topv-nav: Unlocking the top-view spatial reasoning potential of mllm for zero-shot object navigation. *arXiv preprint arXiv:2411.16425*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464.

Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. 2021. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. Ieee.

A Prompts Format

A.1 Prompt used for DeepSeek-V3.2 to annotate the outline of the required assistance

The Prompt used for DeepSeek-V3.2 to annotate the outline of the required assistance is shown in Figure 2.

A.2 Test Prompt for Cascade Models

The Test Prompt for the Cascade Models is shown in Figure 3

You are a robot that proactively engages in conversations with humans based on the scenes you see and sounds you hear, and interacts with the environment to support the conversation process. Your role is to take the initiative to assist the staff in the scene where you are located with their work or activities. For example, helping residents perform life-related operations at home, or assisting teachers in conducting teaching-related activities in a classroom.

You are currently in a picture scene and can hear a segment of background audio at the same time. It should be noted that the description of this scene is from your first-person perspective, and your next plans should be based on this perspective.

I will provide you with a description of the picture scene, a description of the audio, and the logical relationship between the picture and the audio. Please, based on your combined understanding of this picture scene and audio, determine whether the current situation provides you with clear information indicating what to do next, rather than an ambiguous situation where there are multiple possible choices. If there is no clear information, output "no"; if there is clear information, specify both the initiative action you will take next and the purpose of doing so. You also need to output the general content of what you are going to say. It should be noted that the output content must be consistent with the picture scene, and you must not output things or operations that do not exist in the picture. You may choose to output multiple things to do, with each separated by a period. Do not output redundant information.

The description of the picture scene is: ***{ picture_description }***.
The description of the background audio is: ***{ audio_description }***.
The logical relationship between the picture and the audio is: ***{ logical_relationship }***.

Figure 2: Prompt used for DeepSeek-V3.2 to annotate the outline of the required assistance. The highlighted parts represent, in each piece of data, the textual descriptions of pictures, the textual descriptions of audio, and the logical relationships between pictures and audio.

A.3 Test Prompt for End-to-End Models

The Test Prompt for End-to-End Models is shown in Figure 4.

B Details of the PEAP dataset

B.1 Scenes included in the PEAP dataset

In the PEAP task, we collect a total of 122 scene data samples and divide these scenes into 10 categories. Table 10 details the specific names of these 10 scene categories, the number of scenes included in each category, as well as the distribution of corresponding data volumes.

Below, we list in detail the specific scenes included in each category in Figure 5.

B.2 Display of data sample

To facilitate readers' understanding of the data format used in this study, a complete data sample is presented below. Each data entry includes the visual scene, the corresponding audio in that scene, a description of the relationship between the visual scene and the audio, as well as the manually annotated target response that the model should generate.

The data example is shown in Figure 6.

In the above data sample, the models are required to generate a response based on the input visual scene and audio information. We score the model according to the matching degree between the generated response and the annotated target response.

B.3 Case Study

This section is our Case Study on specific data. The Figure 7 shows the model's response to the data presented in B.2.

In the data sample presented, the sound of a siren emanates from the fire station. A qualified model should actively recognize this auditory cue, infer the occurrence of a potential emergency, and then initiate preparatory actions to facilitate the dispatch of fire trucks. In this case study, the text highlighted in green indicates correct responses generated by the model. For instance, the outputs of Llama3-

Category	Number of Scenes	Data Quantity	Test Set
Commercial & Shopping	24	5497	550
Education & Culture	15	1839	181
Entertainment & Leisure	9	1051	105
Medical & Healthcare	8	727	73
Office & Industrial	14	2263	229
Outdoor & Semi-open Spaces	7	829	85
Public Safety & Services	6	481	47
Residential & Private Spaces	14	3723	375
Specialized Functional Spaces	4	738	73
Transportation & Public Infrastructure	21	2815	280
Total	**122**	**19963**	**1998**

Table 10: Scenes included in the PEAP dataset. This table introduces data on the 10 main categories of scenes included in PEAP, including the number of sub-scenes in each category, the total data volume, and the data volume in the test set.

8B and Qwen3-32B are deemed accurate, as they demonstrate precise event recognition and appropriate assistance provision. By contrast, the text highlighted in grey denotes erroneous responses, such as those from MiniCPM, Qwen3-Omni and VITA. These flawed outputs either fail to recognize the emergency implied by the siren or propose irrelevant actions, such as offering to introduce the fire station.

A deeper analysis reveals that the performance discrepancy across models stems from the gap in their capabilities to model semantic correlations between audio-visual modalities and their level of prior knowledge about scenario-specific contexts. Llama3-8B and Qwen3-32B can accurately correlate siren sounds with the semantics of emergency dispatch scenarios, and derive reasonable actions by integrating the contextual attributes of fire stations. In contrast, MiniCPM, Qwen3-Omni and VITA fail to establish such cross-modal correlations, or lack sufficient prior knowledge of emergency scenarios. This limits them to capturing only the surface-level information of individual modalities, preventing them from performing deep reasoning from "audio signals" to "scenario requirements" and ultimately leading to irrelevant or erroneous responses.

B.4 Redandant Output Example

We choose precision and recall as our evaluation metrics because precision reflects the proportion of model-predicted actions that match the ground truth. Low precision indicates that the model outputs a large number of task-irrelevant actions, meaning significant action redundancy exists. Recall reflects the proportion of ground-truth actions that are successfully predicted by the model. When the model's recall rate is significantly higher than its precision, it indicates that the model tends to boost recall by guessing numerous irrelevant actions, which in turn lowers precision.

We analyze this in combination with concrete examples. For instance, in the scenario where an alarm is triggered at a fire station, the model generates the output shown in Figure 8. The highlighted action is meaningless in an emergency context. Although the embodied agent responded to the emergency situation, this action is redundant because it does not directly target solving the problems arising in the scene.

B.5 Different Action Sequences For One Scene

The answers annotated by large language models represent the general content of the appropriate assistance directions corresponding to each scenario combined with audio. They are descriptions of assistance directions rather than concrete action sequences. Therefore, multiple action sequences generated by different models can both meet the requirements and be accepted after being judged by the decision model. This ensures that our task can, to a certain extent, adapt to situations where different responses occur in the same real-world scenario, as long as they remain consistent with the direction of the annotated answers. An example is provided below in Figure 9.

You are a robot that can proactively engage in conversations with humans based on the scenes it observes and the sounds it hears. It interacts with the environment to support the conversation process, and its role is to actively assist the staff in the scene where it is located with their work or activities. For example, when at home, it helps residents perform life-related operations; when in a classroom, it assists teachers in carrying out teaching-related activities.

Currently, you are in a scene depicted in an image and can simultaneously hear the sound of a background audio clip. It is important to note that the description of this scene is from your first-person perspective, and your next steps should be based on this perspective.

I will provide you with a description of the image scene, a description of the audio, as well as the range of actions you can take and corresponding examples. Please, based on your combined understanding of the image scene and the audio, carry out the subsequent consecutive steps of action in accordance with the rules I have set.

The description of the image scene is: ***{ picture_description }.***

The description of the background audio is: ***{ audio_description }.***

The range of actions you can take falls into three main categories.

The first category is [Movement], which includes actions that change your own position and direction, such as walking, moving, and turning. These actions are mainly activities performed from your own perspective.

The second category is [Manipulation], which includes actions that occur when interacting with objects, such as grasping, placing, pushing, pulling, and rotating. The difference between [Manipulation] and [Movement] is that [Manipulation] requires objects other than the robot itself, such as taking a towel, picking up a piece of garbage, and pouring a glass of water.

The third category is [Conversation], which includes all behaviors related to dialogue, including asking questions (Ask), answering questions (Answer), and proactively bringing up a topic (Raise a Topic). This category indicates the dialogue strategy for the current step.

When outputting your next action or statement, first output the label of the action or statement (starting with the major category label followed by the minor category label), and then output the specific content.

For example:

[Conversation][Raise a Topic] That drilling sound seems quite close - perhaps they're doing renovations in the adjacent unit?

[Movement][Turn] Rotate toward the direction of the sound to better assess its source.

[Manipulation][Grab] Pick up the electric kettle from the countertop.

[Conversation][Ask] Would you like me to prepare some tea or coffee while we wait? The hot water might help mask the noise somewhat.

[Movement][Change Position] Move toward the sink to fill the kettle with water.

Print only the dialog itself, with each line at the beginning of the category.Don't print the rest of the dialog, but print a few more lines.

Figure 3: Test Prompt for the Cascade Models. The highlighted parts represent the textual descriptions of pictures and audio, as well as the introduction to the three types of actions.

You are a robot that can proactively engage in conversations with humans based on the scenes it observes and the sounds it hears. It interacts with the environment to support the conversation process, and its role is to actively assist the staff in the scene where it is located with their work or activities. For example, when at home, it helps residents perform life-related operations; when in a classroom, it assists teachers in carrying out teaching-related activities.

Currently, you are in a scene depicted in an image and can simultaneously hear the sound of a background audio clip. It is important to note that the description of this scene is from your first-person perspective, and your next steps should be based on this perspective.

I will provide you with a description of the image scene, a description of the audio, as well as the range of actions you can take and corresponding examples. Please, based on your combined understanding of the image scene and the audio, carry out the subsequent consecutive steps of action in accordance with the rules I have set.

The range of actions you can take falls into three main categories.

The first category is [Movement], which includes actions that change your own position and direction, such as walking, moving, and turning. These actions are mainly activities performed from your own perspective.

The second category is [Manipulation], which includes actions that occur when interacting with objects, such as grasping, placing, pushing, pulling, and rotating. The difference between [Manipulation] and [Movement] is that [Manipulation] requires objects other than the robot itself, such as taking a towel, picking up a piece of garbage, and pouring a glass of water.

The third category is [Conversation], which includes all behaviors related to dialogue, including asking questions (Ask), answering questions (Answer), and proactively bringing up a topic (Raise a Topic). This category indicates the dialogue strategy for the current step.

When outputting your next action or statement, first output the label of the action or statement (starting with the major category label followed by the minor category label), and then output the specific content.

For example:

[Conversation][Raise a Topic] That drilling sound seems quite close - perhaps they're doing renovations in the adjacent unit?

[Movement][Turn] Rotate toward the direction of the sound to better assess its source.

[Manipulation][Grab] Pick up the electric kettle from the countertop.

[Conversation][Ask] Would you like me to prepare some tea or coffee while we wait? The hot water might help mask the noise somewhat.

[Movement][Change Position] Move toward the sink to fill the kettle with water.

Print only the dialog itself, with each line at the beginning of the category.Don't print the rest of the dialog, but print a few more lines.

Figure 4: Test Prompt for End-to-End Models. The highlighted parts represent the introduction to the three types of actions.

Commercial and Shopping: bakery, bar, bazaar, bookstore, buffet, clothing store, deli, fast food restaurant, florist, grocery store, ice cream parlor, jewellery shop, mall, market, plaza, pub, restaurant, restaurant kitchen, restaurant patio, shoe shop, shopfront, supermarket, toy store, video store

Education and Culture: archaeological excavation, auditorium, church inside, classroom, cloister, computer room, concert hall, excavation, kindergarten, kindergarten classroom, library, movie theater, museum, schoolhouse, TV studio

Entertainment and Leisure: amusement park, art studio, ball pit, bowling alley, casino, playground, promenade deck, sandbox, music studio

Medical and Healthcare: dental office, dentist's office, gym, hair salon, hospital, hospital room, operating room, veterinarian's office

Office and Industrial: conference room, construction site, corridor, elevator, engine room, factory, industrial area, wet laboratory, lobby, meeting room, office, office building, staircase, warehouse

Outdoor and Semi-open Spaces: building facade, greenhouse, park, residential neighborhood, skyscraper, street, yard

Public Safety and Services: fire station, gas station, laundromat, prison cell, waiting room, police station

Residential and Private Spaces: basement, bathroom, bedroom, children's room, closet, dining room, game room, hotel room, kitchen, living room, motel, nursery, pantry, wine cellar

Specialized Functional Spaces: firing range, kennel, locker room, indoor pool

Transportation and Public Infrastructure: airplane cabin, airport inside, alley, boardwalk, bridge, bus interior, car interior, crosswalk, driveway, garage, hangar, highway, inside bus, inside subway, parking lot, repair shop, runway, subway, subway station, toll plaza, train station

Figure 5: The specific scenes included in 10 categories. This figure introduces the specific scenes included in each of the 10 categories.

Visual Scene: This photograph depicts ***a vintage red fire truck parked in front of a historic fire station***. The fire truck, with a classic, boxy design and large white-walled tires, features a long, silver metal ladder extending from the top. The truck's front has a black grille and round headlights, and it is adorned with various firefighting equipment, including hoses and a flag on the driver's side. The background reveals an ornate, green-painted fire station with intricate lattice work under a covered porch. ***Two people stand near the truck***: a man in a blue jacket and jeans, and a woman in a blue jacket and white pants. Another red fire truck is partially visible to the right. The overall scene is well-lit, with a clear, daytime atmosphere.

Audio: The sound of ***siren*** is heard.

Relationship between the visual scene and the audio: A siren sound is heard in a fire station.

Annotated target response: I will approach the man and woman near the fire truck to ***alert them to the siren and prepare for a potential emergency response***. I will say, "Excuse me, I hear the siren. We should prepare the fire truck for a potential call."

Example of model response:

[Movement][Turn] Rotate toward the direction of the ambulance siren to monitor its proximity and trajectory.

[Conversation][Ask] Are you both aware of the siren's location? Would you like me to check for any nearby hazards or ***ensure the area is clear for emergency vehicles***?

[Manipulation][Grab] Secure the British flag from the fire truck's side panel ***in case it needs to be relocated for visibility***.

[Movement][Change Position] Step aside to ***create a clear pathway for the fire truck*** if it approaches this location.

Figure 6: Data Example. Each piece of data includes the visual scene, the audio in the scene, the logical relationship between the visual and audio, the annotated standard response, and an example of the model response.

Llama3-8B:

[Conversation][Raise a Topic] Ah, it seems like we have an emergency situation unfolding before our eyes! With the ambulance siren blaring in the background, I'm guessing there might be an accident or a medical emergency nearby. Should we try to locate the source of the sound and offer assistance?

[Conversation][Ask] Is there anything we can do to help, or should we wait for the authorities to arrive?

[Conversation][Answer] I see the man in the blue jacket is already heading towards the sound, perhaps to assess the situation. I'll follow his lead and see if I can be of any assistance.

[Movement][Follow] Move towards the man in the blue jacket to provide support and get a better view of the situation.

Qwen3-32B:

[Movement][Turn] Rotate toward the direction of the fire siren to monitor its proximity and trajectory.

[Conversation][Ask] Are you both aware of the fire siren's location? Would you like me to check for any nearby hazards or ensure the area is clear for emergency vehicles?

[Manipulation][Grab] Secure the British flag from the fire truck's side panel in case it needs to be relocated for visibility.

[Movement][Change Position] Step aside to create a clear pathway for the fire truck if it approaches this location.

MiniCPM:

[Conversation][Raise a Topic] The vintage fire trucks look impressive today.

[Movement][Turn] Turn to face the group of people near the vehicles, ready for interaction.

[Conversation][Ask] Do you need any assistance or information about these historic fire engines?

[Movement][Change Position] Move slightly closer to be more approachable and engaged in conversation.

Qwen3-Omni:

[Conversation][Raise a Topic] That siren is really loud—must be part of the fire truck exhibit demonstration!

[Movement][Turn] Orient toward the two people standing near the vintage fire trucks to engage with them directly.

[Conversation][Ask] Are you part of the museum staff demonstrating the fire truck's functions, or would you like me to help explain how the siren works to visitors?

VITA:

[Conversation][Raise a Topic] The vintage fire truck in the image looks quite impressive. Do you know its history or any stories associated with it?

[Movement][Turn] Turn slightly to get a better view of the fire truck and the people around it.

[Manipulation][Point] Point towards the fire truck to emphasize its presence in the image.

Analysis:

In the data sample presented, the sound of a siren emanates from the fire station. A qualified model should actively recognize this auditory cue, infer the occurrence of a potential emergency, and then initiate preparatory actions to facilitate the dispatch of fire trucks. In this case study, the text highlighted in green indicates correct responses generated by the model. For instance, the outputs of ***Llama3-8B and Qwen3-32B*** are deemed accurate, as they demonstrate ***precise event recognition and appropriate assistance provision***. By contrast, the text highlighted in red denotes erroneous responses, such as those from ***MiniCPM, Qwen3-Omni and VITA***. These flawed outputs either ***fail to recognize the emergency implied by the siren or propose irrelevant actions,*** such as offering to introduce the fire station.

Figure 7: The case study of the data example, the yellow parts are the model names, the green-highlighted parts indicate that the model made correct responses, and the gray-highlighted parts are the incorrect responses.

In the scenario where an alarm sounds in the fire station, the model generated the following sequence of actions :

[Conversation][Ask] Is there anything I can assist with from the fire truck's equipment? I can retrieve the rescue blanket or help organize the gear.

[Movement][Turn] Rotate toward the open side doors to better access the interior equipment.

[Conversation][Comment] The fire truck's paint looks very clean and well-maintained.

[Manipulation][Grab] Pick up the green rescue blanket from the ground to prepare for potential use.

[Conversation][Raise a Topic] The police siren is quite urgent—should I alert nearby personnel or assist in preparing for a possible emergency response?

Within this action sequence, the action ****[Conversation][Comment] The fire truck's paint looks very clean and well-maintained**** is meaningless in an emergency scenario. Although the embodied agent responded to the emergency, this action is redundant because it does not directly address the problems arising in the scene.

Figure 8: A redundant output example. The highlighted action in red represents a redundant action in the model's output.

For example, when hearing ***a distant alarm siren in a fire station***, the following two sets of Action sequences are both acceptable, ***despite their different action orders and representations***.

Action Sequence 1

[Conversation][Raise a Topic] That siren sound means an emergency response is being activated.

[Movement][Turn] Orient my sensors toward the fire trucks to observe the crew's preparation activities.

[Conversation][Ask] Would you like me to verify equipment status or assist with any final checks before departure?

Action Sequence 2

[Conversation][Inform] The siren indicates that an emergency response is now being initiated.

[Conversation][Conversation][Ask] Before departure, would you like me to assist with final inspections or verify the status of the equipment?

[Movement][Turn] Direct my sensors toward the fire trucks to monitor the crew's preparation procedures.

Figure 9: Different action sequences for one scene, both are acceptable.