

Robot Confirmation Generation and Action Planning Using Long-context Q-Former Integrated with Multimodal LLM

Hori, Chiori; Masuyama, Yoshiki; Jain, Siddarth; Corcodel, Radu; Jha, Devesh K.; Romeres, Diego; Le Roux, Jonathan

TR2025-167 December 10, 2025

Abstract

Human-robot collaboration towards a shared goal requires robots to understand human action and interaction with the surrounding environment. This paper focuses on human-robot interaction (HRI) based on human-robot dialogue that relies on the robot action confirmation and action step generation using multimodal scene understanding. The state-of-the-art approach uses multimodal transformers to generate robot action steps aligned with robot action confirmation from a single clip showing a task composed of multiple micro steps. Although actions towards a long-horizon task depend on each other throughout an entire video, the current approaches mainly focus on clip-level processing and do not leverage long-context information. This paper proposes a long-context Q-former incorporating left and right context dependency in full videos. Furthermore, this paper proposes a text-conditioning approach to feed text embeddings directly into the LLM decoder to mitigate the high abstraction of the information in text by Q-former. Experiments with the YouCook2 corpus show that the accuracy of confirmation generation is a major factor in the performance of action planning. Furthermore, we demonstrate that the long-context Q-former improves the confirmation and action planning by integrating VideoLLaMA3.

IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2025

Robot Confirmation Generation and Action Planning Using Long-context Q-Former Integrated with Multimodal LLM

Chiori Hori¹, Yoshiki Masuyama¹,
Siddarth Jain¹, Radu Corcodel¹, Devesh Jha¹, Diego Romeres¹, Jonathan Le Roux¹

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

Abstract—Human-robot collaboration towards a shared goal requires robots to understand human action and interaction with the surrounding environment. This paper focuses on human-robot interaction (HRI) based on human-robot dialogue that relies on the robot action confirmation and action step generation using multimodal scene understanding. The state-of-the-art approach uses multimodal transformers to generate robot action steps aligned with robot action confirmation from a single clip showing a task composed of multiple micro steps. Although actions towards a long-horizon task depend on each other throughout an entire video, the current approaches mainly focus on clip-level processing and do not leverage long-context information. This paper proposes a long-context Q-former incorporating left and right context dependency in full videos. Furthermore, this paper proposes a text-conditioning approach to feed text embeddings directly into the LLM decoder to mitigate the high abstraction of the information in text by Q-former. Experiments with the YouCook2 corpus show that the accuracy of confirmation generation is a major factor in the performance of action planning. Furthermore, we demonstrate that the long-context Q-former improves the confirmation and action planning by integrating VideoLLaMA3.

Index Terms—Human-robot interaction, Robot confirmation generation, Robot action planning, Multimodal scene understanding, Multimodal LLM

I. INTRODUCTION

Human-robot interaction using natural language has the potential to be the most effective way for human-robot collaboration for shared tasks in daily life. Human-to-human collaboration is easily achieved because humans share the required knowledge about tasks and surrounding environments, and can understand other humans’ behaviors. If there are unknown things, humans can use natural language to confirm what to do and how to do it. Such fundamental functions for human-robot collaboration can be built using multimodal scene understanding that enables robots to interpret their environment and interact with humans based on such understanding. This goal lies at the intersection of multiple avenues of research in speech understanding, audio event detection, object and action recognition in computer vision, physical sensing/manipulation for robot control, and natural language generation to interact with humans. Achieving effective human-robot collaboration requires significant advances in the following areas: (1) multimodal scene understanding for human and robot environments, (2) robot action planning based on multimodal understanding

and logical and physical affordance, (3) robot action generation/execution according to the planning, (4) robot action replanning if necessary, and (5) human-robot interaction via dialogue to achieve the goals efficiently.

There are several works on robotic manipulation actions at a high level that have proposed how instructions can be stored and analyzed [1]–[3]. Initial work utilized contrastive learning to learn a reward function to train reinforcement learning (RL) agents [4]. More recently, there have been some works using robot primitives and extraction of cost functions from task videos to enable imitation of demonstrated tasks [5]. There has also been some work on training perception modules on large sets of manipulation data to simplify learning of manipulation tasks [6]. Finally, there has been growing interest in using vision and language for learning diverse robot skills. There are some works on training visual language models using human instruction videos that are well aligned to robot actions to generate action sequences [7], [8]. Then, multimodal language models have been applied to robot action planning using various features, such as audio, visual, speech, and text, to expand knowledge acquisition of the tasks from human demonstration videos. Initially, multimodal scene-understanding-based robot action planning applied an audio-visual transformer trained from cooking videos [9], where the action sequence was translated from human action descriptions.

Recently, neural language models such as Large Language Models (LLMs) have been applied to bridge the gaps between sensing information and abstract-level understanding for robot action planning. This framework allows us to implement an end-to-end approach to build a human-robot collaboration system directly from multimodal scene understanding addressing all the steps (1) to (5). Promising use cases of LLMs were reported in the robotics research field, such as CLIPort [10] and SayCan [11] in creating robotic agents that perform open-vocabulary tasks. PROG PROMPT [12] introduces a programmatic LLM prompt structure that facilitates the generation of plans in diverse environments, robot functionalities, and tasks. LLM-POP [13] targets partially observable task planning, where LLMs expand vocabulary and context considerations, while visual grounding LLMs enhance spatial reasoning capabilities. COWP [14] introduces an LLM-based open-world task planning system for robots. Some works explore using

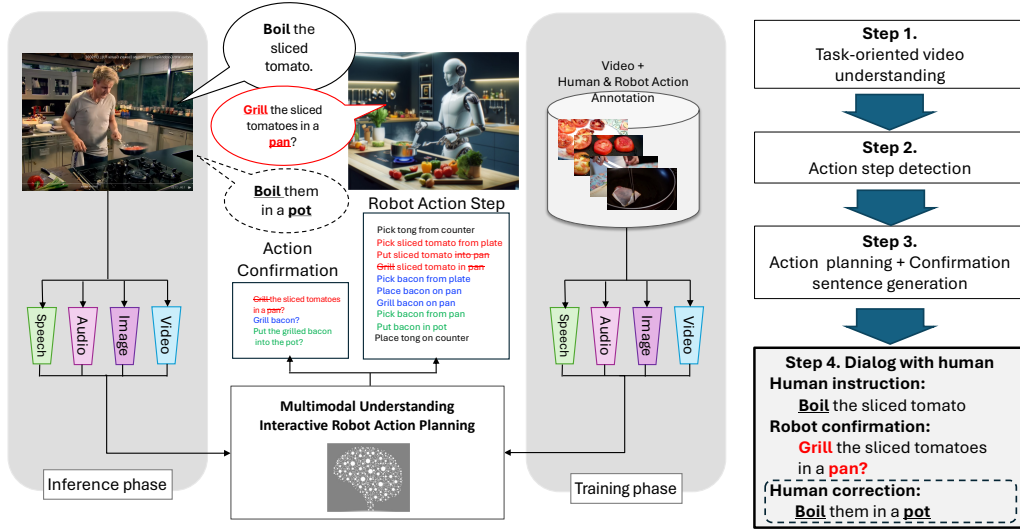


Fig. 1. Robot action confirmation sentences in natural language are simultaneously generated with micro-step action sequences. The multimodal features, such as videos, images, audio, and speech, are extracted from the human demonstration videos, and multimodal LLM models are trained to generate robot action descriptions to confirm whether the action is correct before taking the action. Although a humanoid robot is illustrated in this example, the action steps are designed for commoditized single-arm robots.

LLMs to directly predict a dense sequence of end-effector poses for robot actions with vision models [15]. Another study [16] explores re-prompting strategies to enhance the executability and accuracy of LLM-generated plans but relies strictly on templated prompts. Multimodal scene understanding has significantly advanced with AV-transformers, and has been applied to robot action planning with human action videos [9]. Despite its success, its capability for multimodal reasoning is still limited because it was impractical to cover all combinations of multimodal inputs from the existing videos.

To mitigate the discrepancy issues between the training and inference stages due to the data sparsity in multimodal fusion, BLIP2 [17], narrowing down the semantic space, was applied to train a multimodal large language model (AVBLIP) for robot action planning [18]. Various features are embedded into the semantic space of LLMs by training a Q-former (cross-token transformer), which is an effective method in terms of computational efficiency and information retention compared to conventional fusion methods [19], [20], using both contrastive loss and action generation loss in this approach. The Q-former-based multimodal LLM contributes to enhancing the performance of robot action planning.

However, automatic action planning errors are inevitable because of remaining discrepancies with the training environments and unexpected circumstances. To avoid robots executing incorrect actions, a neural action replanning approach based on human error correction was proposed, where humans intervene to correct the action plans confirmed by robots using natural language before executing the planned actions [21]. In this approach, the robot manipulation was segmented into a skill acquisition phase and a knowledge acquisition phase. Although there have been some other research on language acquisition by robots to find associations between actions, objects, properties, and effects, and to map those associations to language [22], [23], it is unrealistic to train models handling

a huge vocabulary using human demonstration videos through real human-robot interaction. To transfer (2) action planning trained from human action demonstration videos to (3) robot action generation/execution, a robot simulator was applied to generate and optimize executable robot actions [24]. In the next step, we build a system that can handle the real world beyond simulators [25]–[27].

In this research trend, this paper focuses on human-robot interaction (HRI) based on human-robot dialogue that relies on the robot action confirmation and action step generation. The state-of-the-art approach leveraging AVBLIP generates robot action steps aligned with robot action confirmation from a single clip showing a task composed of multiple micro steps. Although action sequence dependency can be captured through an entire video to achieve the goal of a long-horizon task, the current single-clip-based approaches do not apply such long context information. This paper proposes a long-context Q-former incorporating left and right context dependency in full videos. Furthermore, the conventional AVBLIP feeds text embeddings into multimodal Q-former and could unexpectedly be degraded due to the high abstraction of the language information in the multimodal semantic space, thereby losing rich information represented in concrete words. To retain the original rich language information, this paper proposes a text-conditioning approach to feed text embeddings directly into the LLM decoder that generates robot action confirmation and robot action steps. Finally, we show that the long-context Q-former improves the confirmation and action planning by integrating VideoLLaMA3 [28].

The main contributions of this work consist of (a) proposing a long-context Q-former for robot action confirmation sentence generation, (b) introducing a text conditioning that feeds text embedding vectors to an LLM decoder to retain original language information, (c) demonstrating the effectiveness of the proposed approaches for robot confirmation sentence and

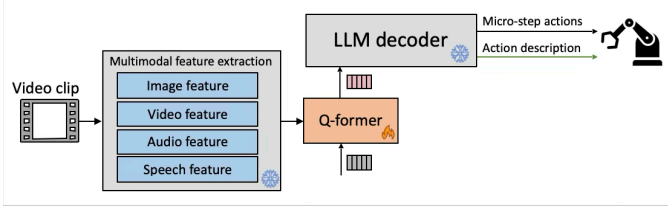


Fig. 2. Q-former-based confirmation sentence generation and action planning model [21]. AVBLIP-based action generation with a Q-former. The generated embeddings are fed to the LLM decoder.

robot action sequence generation in the cooking domain, and (d) examining the contribution of VideoLLaMA.

II. ACTION PLANNING DATA

a) Robot Action Description: YouCook2¹ is already annotated with human instructions in natural language to describe human cooking action steps, as introduced in [29]. The cooking action steps for each video are annotated with start and end time stamps and English description. An example of the description is “Grill the tomatoes in a pan and then put them on a plate” starting from 00:21 and ending at 00:52.

b) Robot Micro-step Action: In the work of [21], the human instructions were translated into a micro-step action sequence such that a single-arm robot could achieve the same goals as humans demonstrated actions, as illustrated in Fig. 1. Although a humanoid robot is illustrated in this example, action steps are designed for commoditized single-arm robots. The micro-step action sequences for single-arm robot action are represented by “single-arm action”, “target object”, “preposition”, and “place”, to achieve the same actions as humans. We borrowed the same data conditions used in [21]. Single-arm actions were selected from the following 12 candidates: Open, Close, Pick, Place, Pour, Stir, TurnOn, TurnOff, Wipe, Cut, Scoop, Squeeze. The target objects were selected as one of the nouns in the human action instruction as much as possible.

III. ROBOT ACTION GENERATION USING AVBLIP

This work employs AVBLIP [18], where BLIP-2 [17], a vision-language pre-training method, is extended to handle multimodal features. BLIP-2 bootstraps from a frozen image encoder and a frozen LLM, where a Querying Transformer (Q-former) [30] is trained to bridge the gap between the vision and text modalities. The image encoder in BLIP-2 is replaced with audio-visual encoders for video, audio, and text feature sequences in AVBLIP. Figure 2 illustrates an architecture of AVBLIP consisting of a Q-former and an LLM decoder. The Q-former is trained to extract a fixed number of embeddings from a multimodal encoder that outputs sequences with different lengths. The self-attention layers are shared between two transformer submodules: (1) a multimodal transformer that interacts with the frozen audio-visual encoders and (2) a text transformer that works as a text encoder and a text decoder. A set of learnable query embeddings is input to the multimodal transformer. The queries interact with each other through self-attention layers and interact with audio-visual features through

cross-attention layers. The queries can additionally interact with the text through the same cross-attention layers. Finally, the queries are converted to an output feature.

The LLM Decoder generates a sequence of micro-step actions for a single-arm robot from multimodal features aligned to language features obtained by the Q-former. The LLM Decoder is constructed with a frozen LLM and a feed-forward layer. By using the LLM as a decoder, it leverages the LLM’s inference capabilities when generating action sequences. In this study, we use OPT-2.7B [31] as the LLM.

The training of AVBLIP consists of two stages: (1) vision-language representation learning with frozen multimodal encoders and (2) vision-to-language generative learning with a frozen LLM. In the second stage, we connect the Q-former to the frozen LLM Decoder and perform multimodal action sequence generation. As shown in Fig. 2, the extracted multimodal features from the video clip are converted to token embeddings using the Q-former. The embedding vectors are then projected to the LLM embedding space using a fully-connected layer. Then, the LLM Decoder generates action sequences and descriptions (confirmation sentences) for the given video clip. We use the cross-entropy loss function for the ground-truth sequences in this stage.

IV. LONG-CONTEX Q-FORMER

The AVBLIP model in Fig. 2 can generate action sequences and confirmation sentences for short video clips of around 10-20 seconds. However, these clips are part of a longer instruction video, consisting of 5-10 sequential clips that aim at a single goal, e.g., “cooking meatloaf”. Therefore, incorporating contextual information from the previous and succeeding video clips is a promising extension of the model to generate more accurate sequences, because the video clips are interdependent: for example, some items cooked in a clip may be used in the next clip.

Figure 3 shows an extended AVBLIP architecture based on a long-context Q-former, which contains two Q-former modules, one processes multi-modal features from the current clip as originally used and the other processes the features from the surrounding clips to utilize them to enhance the output for the current clip, where the surrounding clips could contain the current one. Both Q-formers output token embeddings, which are then combined using a transformer encoder. Finally, the LLM decoder receives the combined embeddings to generate micro-step actions for the robot and action description (as a confirmation) for the human. The two Q-formers and the transformer encoder are jointly trained in the same manner as the original AVBLIP framework described in Section III.

V. TEXT CONDITIONING

As demonstrated in [18], [21], the AVBLIP model can generate both micro-step action sequences and action descriptions by feeding query token embeddings to a pre-trained LLM, where the LLM will not further be trained or finetuned. Thus, the Q-former is trained to generate token embeddings that convey the semantic information of the video clip as

¹<http://youcook2.eecs.umich.edu/>

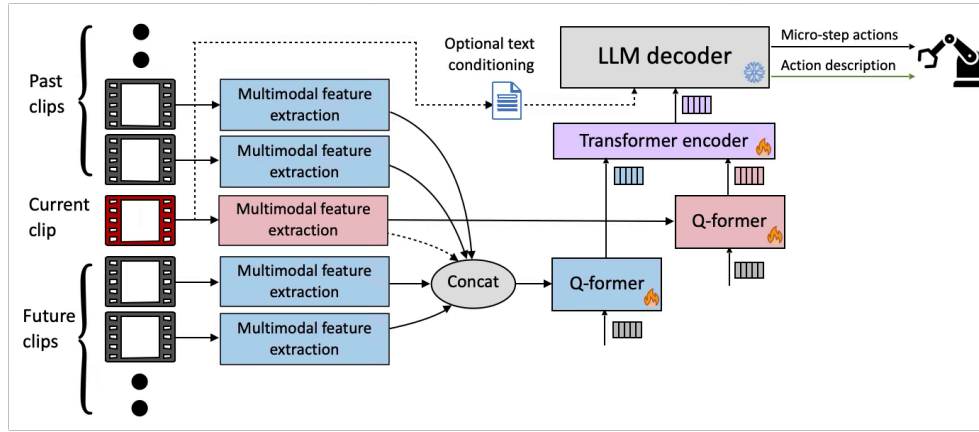


Fig. 3. Long-context Q-former-based confirmation sentence generation and action planning model. AVBLIP-based action generation with two Q-former modules, one generates token embeddings from the current video clip and the other generates the embeddings from the surrounding video clips. The generated embeddings are combined with a transformer encoder and fed to the LLM decoder.

well as instructions to the LLM to generate a micro-step action sequence or an action description. Although the set of embeddings is a compact and effective representation, it may not retain the detailed information necessary for the LLM to generate fine-grained action sequences.

In principle, directly feeding multi-modal features and a natural language instruction to the LLM could be more effective in generating accurate action sequences. However, it is challenging to fine-tune the LLM to understand the multimodal features using limited training data and further to manually create prompts that generate specialized sequences, such as robot micro-step actions. In other words, the Q-former-based approach remains beneficial for obtaining suitable token embeddings, which serve as prompts for the pre-trained LLM. However, there is a concern that precise semantic information could be lost. In particular, text information, such as speech subtitles obtained by automatic speech recognition, often contains exact keywords indicating specific ingredients and utensils. Therefore, it is reasonable that only the text information is fed to the LLM directly with minimal information loss.

This work applies text conditioning to the LLM by prepending the text information to the Q-former token embeddings as in Fig. 3, where the text is tokenized and embedded by the same LLM. The text can be subtitles in the video and/or automatic video descriptions generated by an external multimodal LLM such as VideoLLaMA [28], [32], [33].

VI. EXPERIMENTS

A. Setup

Our proposed methods were tested using YouCook2 consisting of cooking action video clips aligned with human action instruction in natural language. Table I shows the data statistics. In this work, we used the validation set for evaluation since the test set was not publicly available. We borrowed micro-step action sequences [21]. The action set contains 2,790 unique phrases of 195 verbs, 2,229 objects, 33 prepositions, and 1002 places.

Multimodal features such as video, image, and audio were extracted using Omnivore [34], Contrastive Language-Image

TABLE I
YouCookII [29]: “CLIP [SEC]” SHOW THE AVERAGE CLIP DURATION.

| | #Videos | #Clips | Clip [sec] |
|------------|---------|--------|------------|
| Training | 1173 | 8743 | 19.7 |
| Validation | 416 | 3117 | 19.8 |

Pre-Training (CLIP) [35], and Audio Spectrogram Transformer (AST) [36], respectively. The image and video features were concatenated and projected to a single video feature sequence before feeding them to the encoder. If a subtitle was available in the video, text features were extracted by Glove word embedding [37]. Otherwise, we fed an embedding vector for the <unk> label. The numbers of dimensions of the audio, visual, and text features are 768, 1024, and 300, respectively.

We initialized the Q-former with the pre-trained weights of BERTbase [38], while the cross-attention layers were randomly initialized. We set the number of dimensions of the hidden layers to 768, which results in 188M parameters in total. In the experiments, we used 32 queries, where each query has a 768-dimensional vector, which equals the hidden dimension of the Q-former. For the long-context Q-former, we used 32 queries for each of the Q-former modules. The transformer encoder to combine the contextual and current token embeddings had two transformer blocks, each of which had a self-attention layer and a full-connected layer with 768-dimensional hidden activations. These Q-formers and the transformer were jointly trained from scratch without any shared parameters. For the text conditioning, speech subtitles annotated by YouTube and video descriptions obtained from VideoLLaMA3 [28] were fed into the LLM decoder.

The performance was evaluated using the BLEU-2 and METEOR scores computed between the generated and ground-truth sequences used in the robotics field [7], [8]. We applied cross-validation, where one half of the validation set was used to select the best-epoch model, and the other half was used to measure model performance. Thus, the shown scores are the average over the two subsets.

TABLE II

EXAMPLES OF ROBOT ACTION CONFIRMATION SENTENCE AND ROBOT ACTION STEPS.

OBJECT-LEVEL ERRORS ARE HIGHLIGHTED IN RED. CORRECTED OBJECT NAMES ARE HIGHLIGHTED IN BLUE. IF THE CORRECTED OBJECTS ARE SUPPORTED BY THE SUBTITLE, THE VIDEO LLaMA3 DESCRIPTION, AND THE CONTEXT, THEY ARE ALSO HIGHLIGHTED IN BLUE.

| Video clip Id | Source | Action sequence | Action description (confirmation) |
|---------------|-------------|---|---|
| sjh57ujp52M,6 | Reference | pick fish, place fish on towel | remove the fish and drain the oil |
| | Baseline | pick spring roll , place spring roll in oil | remove the patties from the oil and drain on paper towels |
| | Proposed | pick fish , place fish on paper towel | remove the fish from the pan and drain on the paper towels |
| | Subtitle | is take them off and I'm just going to place them on a little bit of paper tissue just to let them drain off so when you serve them up they're not all oily look at that | |
| Z5bpo2sBsl8,2 | VideoLLaMA3 | The video showcases the process of frying fish in a pan. The fish is initially placed into hot oil, where it sizzles and turns golden brown. Once cooked to perfection, the fish is lifted out with a spatula and transferred onto paper towels to drain excess oil. This action takes place on a stove top, indicating that the cooking is taking place at home or in a small kitchen setting. | |
| | Reference | pick knife from counter, cut cabbage with knife, place knife on counter | continue chopping the cabbage |
| | Baseline | pick potato from plate, place potato in bowl, pick bowl from counter, place bowl on counter, pick knife from counter, cut potato in bowl, place knife on counter | grate some parmesan cheese on top of the salad |
| | Proposed | pick cabbage from plate, place cabbage on counter, pick knife from counter, cut cabbage with knife, place knife on counter | chop the cabbage |
| | Subtitle | What weight? Probably about 280. Wow. And I've had a... | |
| | VideoLLaMA3 | In the video, a man and a woman are in a kitchen preparing food. The man is talking while standing next to the woman who is cutting cabbage on a wooden board. | |
| | Context | [previous] chop the cabbage / [next] continue chopping the cabbage | |

TABLE III

QUALITY OF ACTION SEQUENCES AND DESCRIPTIONS GENERATED BY LONG-CONTEXT Q-FORMER. THE SECOND COLUMN SHOWS THE CONTEXT DESIGN. FOR EXAMPLE, “-1,0,+1” REPRESENTS THE CONTEXT OF PREVIOUS (-1), CURRENT (0), AND NEXT (+1) CLIPS. “*” DENOTES ALL PREVIOUS OR FOLLOWING CLIPS.

| Model | Context | Action sequence | | Action description | |
|-------------------|---------------|-----------------|--------------|--------------------|--------------|
| | | BLEU-2 | METEOR | BLUE-2 | METEOR |
| Baseline | - | 0.370 | 0.260 | 0.229 | 0.159 |
| | -1,0 | 0.374 | 0.264 | 0.235 | 0.166 |
| | -2,-1,0 | 0.376 | 0.265 | 0.237 | 0.168 |
| Long-Context | *,0 | 0.376 | 0.266 | 0.237 | 0.169 |
| | -1,0,+1 | 0.379 | 0.269 | 0.240 | 0.170 |
| | -2,-1,0,+1,+2 | 0.381 | 0.269 | 0.242 | 0.170 |
| | *,0,* | 0.381 | 0.270 | 0.241 | 0.169 |
| | | | | | |
| Multitask [21] | - | 0.370 | 0.257 | 0.220 | 0.158 |
| ErrorCorrect [21] | - | 0.375 | 0.258 | 0.231 | 0.161 |

B. Results

Table III shows the quality of action sequences and descriptions generated by the long-context Q-former. “Baseline” denotes the Q-former model trained with pairs of multimodal features and their target sequences without contextual information. Unlike the method in [21] that trains a single model to generate both action sequences and descriptions in a multitask manner, our baseline model was trained to generate action sequences and descriptions separately. Although we can use the same technique to achieve slightly better quality, we omitted that process for simplicity. In addition, our baseline model already achieves a quality comparable to the result of multitask training in [21]. This could be because there are some differences in the hyperparameter setting, where we tuned the batch size, the learning rate, and the number of epochs in preliminary experiments.

We can see certain gains in BLEU and METEOR scores by using the long-context Q-former. The three rows after the baseline correspond to the results of left-context expansion, where introducing the previous two clips is sufficient, and we do not have to consider the full left context. The next three rows display the results of introducing both left and right contexts, yielding further gains. Similar to the left-context expansion, it is sufficient to include two previous and two following clips to achieve the best quality. These results

TABLE IV

QUALITY OF ACTION SEQUENCES AND DESCRIPTIONS ENHANCED BY TEXT CONDITIONING.

| Conditioning | Action sequence | | Action description | |
|----------------------------|-----------------|--------------|--------------------|--------------|
| | BLEU-2 | METEOR | BLEU-2 | METEOR |
| Baseline (no conditioning) | 0.370 | 0.260 | 0.229 | 0.159 |
| Speech subtitle | 0.411 | 0.281 | 0.257 | 0.170 |
| VideoLLaMA3 | 0.401 | 0.275 | 0.255 | 0.168 |
| VideoLLaMA3+subtitle | 0.424 | 0.290 | 0.260 | 0.178 |
| Generated description | 0.398 | 0.272 | - | - |
| Ground-truth description | 0.499 | 0.337 | - | - |

demonstrate that the contextual information helps action and description generation and the long-context Q-former can effectively utilize that information. Note that the training time of the long-context Q-former with full-video context (*,0,*) took 8.2 hours on a single A40 GPU, compared to 5.5 hours for the baseline, when performing two-stage training for 30 epochs with batch size 16. The average inference time was 5.8 seconds compared to 4.1 seconds for the baseline. Thus, the computational cost for the long-context model is not crucial.

Table IV shows the quality of action sequences and descriptions generated by the Q-former with text conditioning. First, speech subtitles provided a substantial gain from the baseline in all metrics. This demonstrates that text conditioning of the LLM decoder is very effective for this task. Then, we introduced a different text conditioning with video descriptions generated by the VideoLLaMA 7B model that achieves state-of-the-art performance in multiple video understanding benchmarks [28]. To ask VideoLLaMA3 to generate rich video descriptions for text conditioning, we used a simple prompt “Describe the cooking video in detail.” This approach also improved the quality of the action sequences and descriptions, which are close to those with speech subtitles. Then, we concatenated two text sequences and fed them to the LLM (“VideoLLaMA3+subtitle”), and obtained further improvement. This result shows that rich and precise information is very useful when it is provided as an LLM’s context. Further, we tested the cases that use generated action descriptions and ground-truth descriptions. With the descriptions generated by VideoLLaMA3+subtitle conditioning, the action sequence

TABLE V
QUALITY OF ACTION SEQUENCES AND DESCRIPTIONS GENERATED BY
LONG-CONTEXT Q-FORMER AND TEXT CONDITIONING.

| Context | Text conditioning | Action sequence | | Action description | |
|---------------|----------------------|-----------------|--------------|--------------------|--------------|
| | | BLEU-2 | METEOR | BLEU-2 | METEOR |
| Baseline | - | 0.370 | 0.260 | 0.229 | 0.159 |
| - | VideoLLaMA3+subtitle | 0.424 | 0.290 | 0.260 | 0.178 |
| -2,-1,0 | VideoLLaMA3+subtitle | 0.427 | 0.290 | 0.264 | 0.180 |
| -2,-1,0,+1,+2 | VideoLLaMA3+subtitle | 0.432 | 0.291 | 0.270 | 0.183 |

quality degraded. This is because the generated descriptions were much shorter than those of VideoLLaMA3+subtitle, and a certain level of information could be lost. The ground-truth descriptions were concise, yet they provided accurate descriptions, achieving the highest quality, which is considered the upper bound of the text conditioning approach.

Finally, we combined the long-context Q-former with text conditioning. Table V shows the quality of the generated sequences using the combined approach. The results indicate that the performance gains by long-context Q-former and text conditioning are additive, achieving the best BLEU and METEOR scores. The relative gains of the scores from the baseline are 16.7% and 11.9% for action sequences and 17.9% and 15.1% for action descriptions, respectively.

VII. ANALYSIS AND DISCUSSIONS

Table II shows examples of generated action sequences and descriptions, where the rows contain the reference, the baseline result, the sequence generated by the proposed method, the speech subtitle, and the VideoLLaMA3 description. They also include generated descriptions for the previous and the following clips in the second example. The proposed method used both long-context expansion and text conditioning. In the first example, the errors in the baseline result are corrected by the proposed method, where the word “spring roll” is substituted with the correct word “fish”. This shows that the VideoLLaMA description successfully supported the correct word through text conditioning. In the second example, the object “potato” in the action sequence and “parmesan cheese” in the action description are corrected to “cabbage” using the proposed model. This correction could be made using information from the multimodal features of the previous and following clips, as the model predicted the word. Although the VideoLLaMA description also included “cabbage”, the error could not be fixed when the long-context features were not used. We also have several examples in which speech subtitles support correcting the action sequences and descriptions, but we omit them due to space limitations. As in those examples, the long context and text conditioning provide helpful information to generate accurate action sequences and descriptions.

Finally, we show the impact of text information on the feature side and the prompt side, i.e., text conditioning. Table VI shows the quality of action sequences and descriptions in different feature combinations. The first row shows the baseline result, where we used “base” features consisting of image, video, audio, and subtitle features. The second row shows the result when we removed subtitles from the

TABLE VI
IMPACT OF TEXT INFORMATION ON THE FEATURE SIDE AND THE PROMPT SIDE. “BASE” DENOTES BASIC MULTIMODAL FEATURES WE USED IN THIS WORK, I.E., IMAGE, VIDEO, AUDIO, AND SUBTITLE FEATURES.

| Features | Text conditioning | Action sequence | | Action description | |
|--------------------|-------------------|-----------------|--------|--------------------|--------|
| | | BLEU-2 | METEOR | BLEU-2 | METEOR |
| Base | - | 0.370 | 0.260 | 0.229 | 0.159 |
| Base - subtitle | - | 0.373 | 0.257 | 0.231 | 0.158 |
| Base | Subtitle | 0.411 | 0.281 | 0.257 | 0.170 |
| Base - subtitle | Subtitle | 0.409 | 0.281 | 0.258 | 0.170 |
| Base + VideoLLaMA3 | - | 0.377 | 0.262 | 0.233 | 0.160 |
| Base | VideoLLaMA3 | 0.401 | 0.275 | 0.255 | 0.168 |
| Base + VideoLLaMA3 | VideoLLaMA3 | 0.400 | 0.276 | 0.256 | 0.170 |
| Subtitle only | Subtitle | 0.223 | 0.219 | 0.174 | 0.126 |
| VideoLLaMA3 only | VideoLLaMA3 | 0.353 | 0.252 | 0.211 | 0.151 |

base features. Compared to the baseline, the BLEU and METEOR scores do not change significantly, meaning that the baseline model did not effectively utilize the subtitle feature. The third row indicates that text conditioning with subtitles substantially improves the sequence quality. The fourth row corresponds to the result when we use subtitles on both sides, showing that adding subtitles to the features does not help. We also used VideoLLaMA descriptions as features and/or text conditioning, and obtained similar results to the case of subtitles. The last two rows show the results when we removed the base features, where we used only subtitles or VideoLLaMA descriptions, because the text may already contain rich semantic information of the video, and therefore, we want to check if the base features are still necessary for this task. We can see substantial degradation of the quality, which means that the multimodal features are still important for this task even though a subtitle or rich video description is provided. Furthermore, we found that not all videos have valid subtitles. They include transcripts of unrelated utterances, non-English subtitles, and background music alone, which may not contribute to the performance. Therefore, the multimodal features are essential to compensate for invalid subtitles.

VIII. CONCLUSIONS

This paper proposed a method for robot action sequence and confirmation sentence generation that leverages (a) a long-context Q-former considering left and right context dependency in full videos and (b) a text-conditioned LLM decoder to retain the precise language information. We trained the above proposed models, generating single-arm robot micro-step action sequences and robot action confirmation in natural language using the YouCook2 dataset. To mitigate the sparseness of the human action descriptions, we leveraged the video descriptions obtained from VideoLLaMA3. Experimental results show that our proposed long-context Q-former outperformed the baseline model for all metrics. We confirmed that the combination of all multimodal features and text-conditioning performed the best. Future work includes (1) multitask training of robot action steps and action descriptions, (2) comparison with other multimodal fusion approaches such as simple sequence concatenation or pooling+MLP, (3) evaluation for tasks other than cooking, and (4) evaluation using a simulator or a real robot.

REFERENCES

- [1] M. Tenorth, J. Ziegler, and M. Beetz, "Automated alignment of specifications of everyday manipulation tasks," in *Proc. IROS*, 2013, pp. 5923–5928.
- [2] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos, "A cognitive system for understanding human manipulation actions," *Advances in Cognitive Systems*, vol. 3, pp. 67–86, 2014.
- [3] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web," in *Proc. AAAI*, 2015.
- [4] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *Proc. ICRA*, 2018.
- [5] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," in *Proc. RSS*, 2022.
- [6] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.
- [7] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating videos to commands for robotic manipulation with deep recurrent neural networks," in *Proc. ICRA*, 2018.
- [8] X. Xu, K. Qian, B. Zhou, S. Chen, and Y. Li, "Two-stream 2D/3D residual networks for learning robot manipulations from human demonstration videos," in *Proc. ICRA*, 2021.
- [9] C. Hori, P. Peng, D. Harwath, X. Liu, K. Ota, S. Jain, R. Corcodel, D. Jha, D. Romeres, and J. Le Roux, "Style-transfer based Speech and Audio-visual Scene understanding for Robot Action Sequence Acquisition from Videos," in *Proc. Interspeech*, 2023, pp. 4663–4667.
- [10] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proc. CoRL*, 2021.
- [11] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.
- [12] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *Proc. ICRA*, 2023.
- [13] L. Sun, D. K. Jha, C. Hori, S. Jain, R. Corcodel, X. Zhu, M. Tomizuka, and D. Romeres, "Interactive planning using large language models for partially observable robotic tasks," in *Proc. ICRA*, 2024.
- [14] Y. Ding, X. Zhang, S. Amiri, N. Cao, H. Yang, A. Kaminski, C. Esselink, and S. Zhang, "Integrating action knowledge and LLMs for task planning and situation handling in open worlds," *Autonomous Robots*, vol. 47, no. 8, pp. 981–997, 2023.
- [15] T. Kwon, N. Di Palo, and E. Johns, "Language models as zero-shot trajectory generators," *IEEE Robotics and Automation Letters*, 2024.
- [16] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, "Planning with large language models via corrective re-prompting," in *NeurIPS Foundation Models for Decision Making Workshop*, 2022.
- [17] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. ICML*, 2023.
- [18] M. Kambara, K. Sugiura, S. Khurana, K. Ota, S. Jain, R. Corcodel, D. Jha, D. Romeres, J. Le Roux, and C. Hori, "Human action understanding-based robot planning using multimodal LLM," in *Proc. ICRA Workshop for "Cooking Robotics: Perception and motion planning"*, 2024.
- [19] J. An, J. Lee, J. Lee, and Y. Son, "Towards llm-centric multimodal fusion: A survey on integration strategies and techniques," 2025. [Online]. Available: <https://arxiv.org/abs/2506.04788>
- [20] S. N. Wadekar, A. Chaurasia, A. Chadha, and E. Culurciello, "The evolution of multimodal model architectures," 2024. [Online]. Available: <https://arxiv.org/abs/2405.17927>
- [21] C. Hori, M. Kambara, K. Sugiura, K. Ota, S. Khurana, S. Jain, R. Corcodel, D. Jha, D. Romeres, and J. Le Roux, "Interactive robot action replanning using multimodal llm trained from human demonstration videos," in *Proc. ICASSP*, 2025.
- [22] G. Saponaro, L. Jamone, A. Bernardino, and G. Salvi, "Beyond the self: Using grounded affordances to interpret and describe others' actions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 209–221, 2020.
- [23] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception–action association," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 660–671, 2012.
- [24] K. Lu, C. Ma, C. Hori, and D. Romeres, "Kitchenvla: Iterative vision-language corrections for robotic execution of human tasks," in *Proc. ICRA Workshop for "Safely Leveraging Vision-Language Foundation Models in Robotics : Challenges and Opportunities"*, 2025.
- [25] S. Tao *et al.*, "Maniskill3: GPU parallelized simulation and rendering for generalizable embodied AI," in *Proc. ICRA Workshop for "7th Robot Learning Workshop: Towards Robots with Human-Level Abilities"*, 2025.
- [26] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne, A. Narayan, Y. L. Tan, G. Wang, Q. Wang, J. Xiang, Y. Xu, S. Ye, J. Kautz, F. Huang, Y. Zhu, and L. Fan, "Flare: Robot learning with implicit world modeling," *arXiv preprint arXiv:2505.15659*, 2025.
- [27] Y. Li, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "WorldEval: World model as real-world robot policies evaluator," *arXiv preprint arXiv:2505.19017*, 2025.
- [28] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, "VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.
- [29] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. AAAI*, 2018.
- [30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.
- [31] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan *et al.*, "OPT: Open Pre-trained Transformer Language Models," *arXiv preprint arXiv:2205.01068*, 2022.
- [32] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.
- [33] G. Sun, W. Yu, C. Tang, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, Y. Wang, and C. Zhang, "video-SALMONN: Speech-enhanced audio-visual large language models," in *Proc. ICML*, 2024.
- [34] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A Single Model for Many Visual Modalities," in *Proc. CVPR*, 2022.
- [35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
- [36] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.
- [37] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.
- [38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.