2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

CommCP: Efficient Multi-Agent Coordination via LLM-Based Communication with Conformal Prediction

Xiaopan Zhang*, Zejin Wang*, Zhixu Li, Jianpeng Yao, Jiachen Li‡

***Abstract*— To complete assignments provided by humans in natural language, robots must interpret commands, generate and answer relevant questions for scene understanding, and manipulate target objects. Real-world deployments often require multiple heterogeneous robots with different manipulation capabilities to handle different assignments cooperatively. Beyond the need for specialized manipulation skills, effective information gathering is important in completing these assignments. To address this component of the problem, we formalize the information-gathering process in a fully cooperative setting as an underexplored multi-agent multi-task Embodied Question Answering (MM-EQA) problem, which is a novel extension of canonical Embodied Question Answering (EQA), where effective communication is crucial for coordinating efforts without redundancy. To address this problem, we propose CommCP, a novel LLM-based decentralized communication framework designed for MM-EQA. Our framework employs conformal prediction to calibrate the generated messages, thereby minimizing receiver distractions and enhancing communication reliability. To evaluate our framework, we introduce an MM-EQA benchmark featuring diverse, photo-realistic household scenarios with embodied questions. Experimental results demonstrate that CommCP significantly enhances the task success rate and exploration efficiency over baselines. The experiment videos, code, and dataset are available on our project website: `https://comm-cp.github.io`.**

I. Introduction

Modern service robots are designed to understand human instructions and complete tasks in real-world household environments (e.g., "Turn off the TV if it is currently on," "Bring the red pillow from the living room to the bedroom"). This process involves interpreting natural language commands, generating and answering relevant questions for scene understanding and reasoning (e.g., "Is the TV turned on?" "What is the color of the pillow?"), and manipulating target objects accordingly. A crucial step is answering these questions, a task known as Embodied Question Answering (EQA) [1], which requires robots to efficiently explore the 3D environment from a random starting location and actively gather information until a confident answer can be provided. Prior studies have investigated this in single-agent settings [1]–[5]. In contrast, we envision future households with multiple *heterogeneous* service robots, each with distinct capabilities and non-transferable assignments. While they cannot take over each other's tasks, they can access all generated questions and share observations and interpretations to enhance exploration efficiency. We define this cooperative information-gathering setting as a *multi-agent multi-task EQA (MM-EQA)* problem, a novel challenge that facilitates multi-robot collaboration in real-world scenarios.

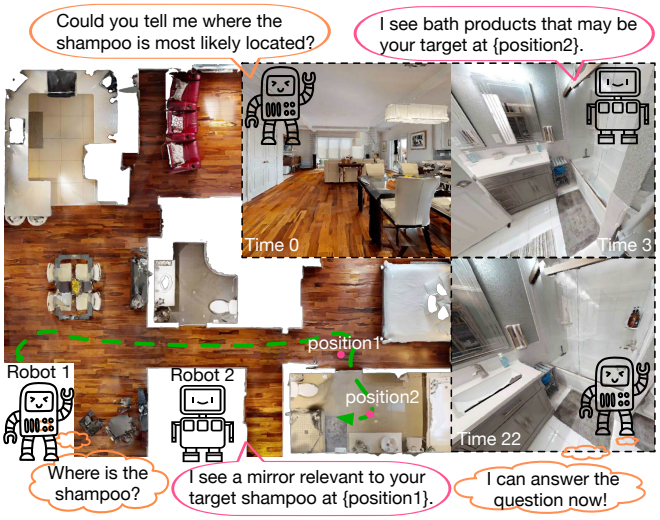

Fig. 1. In a household setting, robots exchange observations and reasoning to collaboratively complete their assigned tasks. Each agent generates confident and goal-directed messages using calibrated outputs from LLMs. The bottom-left image shows a bird's-eye view of Robot 1's navigation path after incorporating information received from Robot 2. The top-right sequence captures both robots' camera views at different timestamps.

While existing single-agent EQA solutions can be adapted to a multi-agent setting by having each robot work independently, this naive approach is inefficient. Communication enables mutual assistance, improving exploration and increasing the likelihood of faster task completion. However, uncalibrated communication could hinder efficiency by sharing irrelevant or misleading information. Therefore, it is critical to ensure that messages are accurate and pertinent to the recipient's tasks. This work tackles the MM-EQA problem by designing a communication framework that enhances multi-agent exploration efficiency and task performance.

Large language Models (LLMs) have shown great potential in solving EQA tasks due to their remarkable ability to understand natural-language queries, reason, and provide answers in natural language [6]. In the context of MM-EQA, natural language is an ideal communication protocol, as LLMs are inherently trained to engage in dialogues. Several LLM-based communication methods have been proposed in other domains [7], [8], but these cannot be directly adapted to our MM-EQA setting. Additionally, LLMs often produce miscalibrated and overconfident outputs [9], which can result in irrelevant or misleading information. This can hinder cooperation efficiency, as agents may share inaccurate data,

*Equal contribution ‡Corresponding author
X. Zhang, Z. Wang, Z. Li, J. Yao, and J. Li are with the Trustworthy Autonomous Systems Laboratory at the University of California, Riverside, CA, USA. {xzhan006, jiachen.li}@ucr.edu.

reducing overall exploration effectiveness [10].

Our work tackles this challenge and develops an LLM-based communication framework for MM-EQA. Our key insight is that *an agent should only communicate information it confidently deems relevant to its partner agents' tasks* (see Fig. 1). We propose CommCP, a novel decentralized LLM-based communication framework that employs conformal prediction (CP) [11], [12] to calibrate the confidence of LLM's outputs. Our framework ensures that the outputs generated by LLMs are more reliable and reduces the negative impact of irrelevant or misleading information to partner agents, ultimately enhancing the overall task performance and efficiency of the multi-agent system. To evaluate our proposed framework, we create a novel MM-EQA benchmark based on realistic scenarios and the Habitat-Matterport 3D (HM3D) dataset [13]. The experimental results show that our approach enhances the task success rate and shortens completion time by a large margin over baselines.

The main contributions of this paper are as follows:

- We formulate the information-gathering process of completing assignments provided in natural language as a novel multi-agent multi-task embodied question answering (MM-EQA) problem, where multiple robots work as a team, handling EQA tasks in a shared environment and communicating to exchange information or answers.
- We propose CommCP, a novel LLM-based decentralized communication framework for MM-EQA, where conformal prediction is employed to calibrate the generated messages to reduce distractions to other agents and improve communication reliability and efficiency.
- We create a novel MM-EQA benchmark with photo-realistic scenarios from the HM3D dataset to validate the effectiveness of the proposed framework. This benchmark is released to facilitate future studies.

II. Related Work

A. LLM-based Decentralized Multi-Agent Cooperation

LLM-based multi-agent cooperation has gained increasing attention recently [14], [15], with various systems developed for multi-agent tasks [7], [16]–[20]. Unlike single-agent or centralized systems, decentralized cooperative systems involve peer-to-peer communication, where agents interact directly, an architecture common in world simulation applications [21], [22]. In these systems, communication typically takes the form of natural language text generated by LLMs, with content varying by application, such as sharing environmental observations, coordinating actions, or reallocating tasks. However, the effectiveness of LLM-generated communication remains underexplored. As noted in [6], decentralized communication often incurs costs such as bandwidth limitations or delays. Thus, agents must communicate efficiently and avoid unnecessary or redundant messages. Current approaches lack mechanisms to assess communication quality and rely solely on raw LLM outputs, leading to inefficiencies, especially when agents act on incomplete or uncertain information.

B. Conformal Prediction and Calibration

Recent research has highlighted the miscalibration issue in LLMs, where models may exhibit overconfidence or underconfidence in their text outputs. This presents a huge challenge as foundation models are applied to embodied tasks where agents may have miscalibrated confidence in their decisions. Previous work [12], [23] has employed conformal prediction [11] to formally quantify an LLM's uncertainty in a robot planning context, which ensures that the robot's plans are executed with calibrated confidence. Explore until Confident [2] extends this approach by applying multi-step conformal prediction in EQA tasks to determine when the VLM is sufficiently confident when a visual language model (VLM) is sufficiently confident to stop exploration. To our best knowledge, we are the first to employ conformal prediction to enhance multi-agent communication through calibrating confidence during collaborative exploration, which is a setting not addressed by prior work.

III. Problem Formulation

Consider a scenario where N_a robots are deployed in a 3D scene with multiple different assignments, each starting from an initial pose g_0^i, and aiming to answer the questions $q_{1:N_q}^i$ related to its assignments. The objective is to maximize the success rate while minimizing the exploration time, with all answers required within a time horizon T_{max}. Each robot knows all questions, including those assigned to others. They can communicate via natural language messages, denoted ζ^i, to exchange information.

Each robot $i \in N_a$ is equipped with cameras that, at each time step t, can provide the robot with an RGB image $I_{c,t}^i$ and a depth image $I_{d,t}^i$ of the local scene as observations. The pose (2D position and orientation) of each robot at each time step is denoted as g_t^i, with the poses of all robots collected into a set $G_t = \{g_t^i \mid i = 1, ..., N_a\}$. Each robot is equipped with a collision-free planner π to navigate. Given the current pose g_t^i and a target position, the planner π determines the next feasible pose g_{t+1}^i, with a low-level controller transporting the robot to the planned pose at $t+1$.

In this case, a multi-robot multi-task Embodied Question Answering (MM-EQA) problem is defined with a tuple $\xi := (E, G_0, T_{\text{max}}, Q, Y)$, where E is the 3D scene with dimensions $L \times W \times H$, which is discretized into a voxel map M composed of cubes with a side length of l. L, W, and H representing the length, width, and height of the voxel map M; $G_0 = \{g_0^i \mid i = 1, ..., N_a\}$ is a set of initial poses of the robots, and T_{max} is the maximum time horizon allowed for the robots to explore the scene and complete the task. Each robot i is assigned with N_q questions to answer. The set $Q = \{q_j^i \mid i = 1, ..., N_{\text{a}}, j = 1, ..., N_{\text{q}}\}$ collects all the questions assigned to the robots, with q_j^i being the j^{th} question assigned to the i^{th} robot. Each question is a multiple-choice question with four choices {'A', 'B', 'C', 'D'}. The ground truth answers are denoted by the set $Y = \{a_j^i \in \{\text{'A', 'B', 'C', 'D'}\} \mid i = 1, ..., N_a, j = 1, ..., N_q\}$.

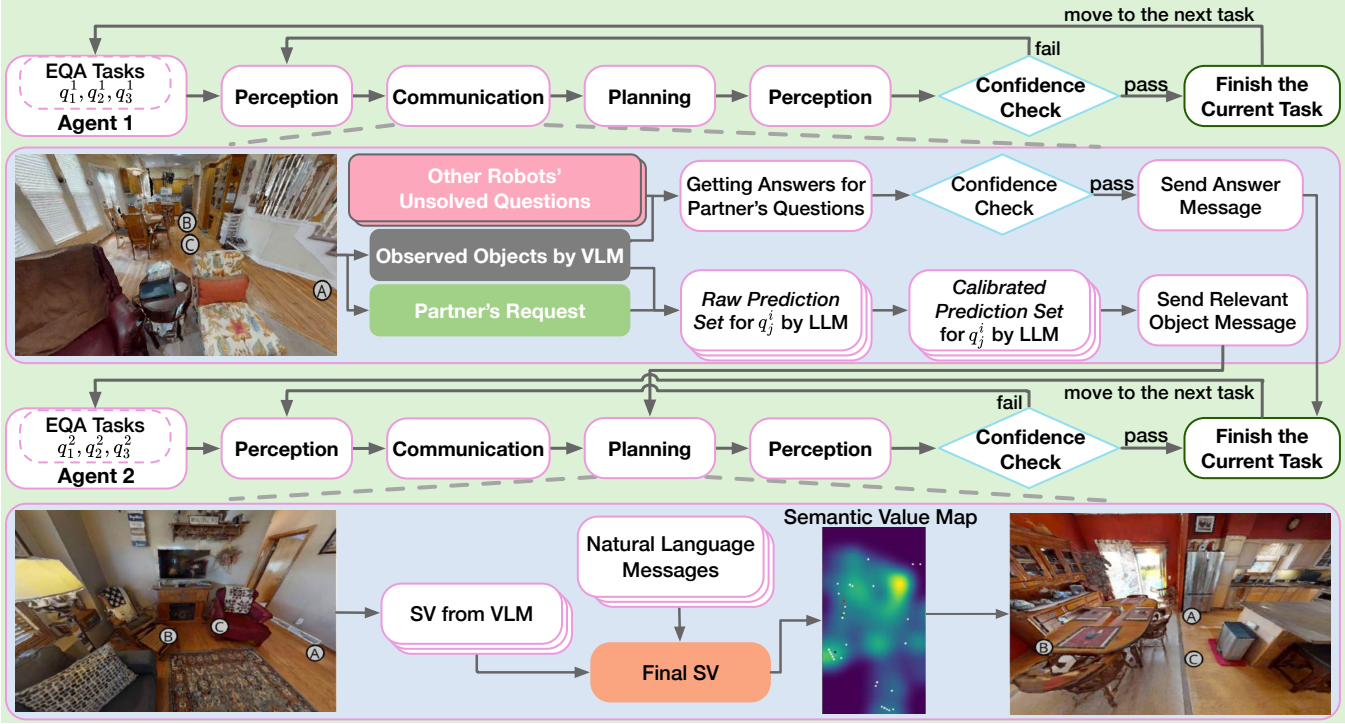

Fig. 2. An overview of our framework shows each robot with a perception module, a communication module, a planning module, and a confidence check module. At each time step, a robot generates local and global semantic values (SV) based on the current view and the communication message from the other robot. It navigates using a 2D weighted semantic value map and handles related object-check requests from other agents. The messages are generated based on the robot's current view, which are calibrated by conformal prediction to enhance relevance.

IV. Method

This section introduces the CommCP framework, which leverages communication to enhance multi-agent exploration for the MM-EQA problem. Building upon the approach presented in [2] for single-agent single-task EQA, our framework further enables communication capabilities to improve task completion efficiency and success rate. Furthermore, our framework can be easily extended to handle more complex human assignments in multi-agent settings, where robots are tasked with executing downstream tasks based on the information they acquire since robots are assigned questions based on their capabilities for the downstream tasks.

The overall framework architecture is illustrated in Fig. 2, which consists of four core modules: perception, communication, planning, and confidence check modules. The planning and confidence check modules are modified from [2]. In the communication module, messages are generated and projected onto the semantic value map in the planning module to guide the robots' exploration strategies. This module also enables a robot to provide answers to other robots' questions when it has sufficient confidence. All notations in this section correspond to time t. For each robot i, the perception module employs a VLM to detect a set of observed objects O^i_{observe} from the RGB image I^i_c with the following prompt:

> *Consider the indoor scenario, analyze the provided image and list all the objects you observe. Provide the name of each object along with its color, separated by a comma.*

The perception module's detected O^i_{observe} are fed into the reasoning process to determine which objects to include in the communication by employing conformal prediction. Each unsolved question is sequentially prompted to the LLM, along with O^i_{observe}, to generate an answer. If the answer passes the confidence check described in Sec. IV.D and the question is assigned to the responding robot, it proceeds. Otherwise, it sends the answer to the responsible robot.

A. LLM-Based Object Relevance Reasoning

In indoor scenarios, objects are typically organized according to patterns of human usage, and LLMs can leverage the general knowledge of these patterns. Thus, if LLM assesses that an object observed in an area is highly relevant to the target object, there is an increased likelihood that the target is nearby. Based on this intuition, we design a communication framework that enables robots to share relevant object information or answers to other robots' questions.

During exploration, each robot $\hat{i}$ sends a request $r^{\hat{i}}_j$ to seek for assistance on its question $q^{\hat{i}}_j$ and provide its target objects $O^{\hat{i}}_{\text{request}}$. The LLM evaluates objects observed by robot i, denoted as O^i_{observe}, and the observed and target objects are labeled as **Observed** and **Request**, respectively. We employ the zero-shot chain-of-thought [24] to prompt the LLM to conduct detailed reasoning before generating a final output. The following is the prompts used for the LLM. Here, the *system prompt* is the instruction for the LLM, and the *user prompt* is the content of the conversation with LLM.

> **System prompt:**
> *As a robot in a house, your partner is looking for {**Request**} and you can inform them about what you have observed.*

User prompt:
*You observe {**Observed**}. Your partner wants to find {**Request**}. Since you and your partner are not in different locations, evaluate whether it is worth it for your partner to travel to your position to find {**Request**}.*
*You have four options: **A** (Yes. Because {**Request**} and {**Observed**} are same, and you might directly find {**Request**}). **B** (Yes. They are highly relevant. This means {**Request**} should be close to {**Observed**}). **C** (No. They are not strongly related). **D** (No. {**Observed**} is a common feature).*
*For option A, consider if {**Observed**} and {**Request**} are the same things. For option B, consider where {**Request**} is most likely to be found. Then evaluate if {**Observed**} is likely to appear in that location as well. For options C and D, since you and your partner are not in the same position, assess if {**Observed**} is worth passing by. Option D implies that {**Observed**} is just a common feature in the house. Now provide your analysis.*

The LLM's output, ***Analysis***, is used to compute probabilities for each option. We employ the following prompt that directs the LLM to select only one option based on ***Analysis***.

System prompt:
You should only output one letter.
User prompt:
*Your observe $O_{t,k}^{observe,i}$ now. Your partner wants to find {**Request**}. Since you are in different locations, you need to assess if it's worthwhile for your partner to come for {**Request**}. You have four options: **A, B, C, D**. Here is your previous analysis: {**Analysis**}. You should only output one letter.*

Each observed object for robot i is assigned an *option-probability pair* $O_{\text{observe},k}^{i} := \{\text{Option}_k, p_k \mid k = 1, ..., N_k\}$, where N_k is the number of observed objects and p_k is the probability of the token corresponding to Option_k generated by the LLM. If **Option A** is returned, it is likely to correspond to the target object $O_{\text{target},j}^{i}$ of request r_j^i. If **Option B** is selected, it is likely a relevant object $O_{\text{relevant},j}^{i}$ related to request r_j^i. Objects with options C and D are disregarded.

B. Message Generation with Conformal Prediction

To mitigate potential miscalibration in LLM-generated lists of targets and relevant objects, which leads to irrelevant information, we employ conformal prediction (CP) [11] to calibrate object lists for message generation. CP provides statistically guaranteed prediction sets that contain the ground truth with a user-specified probability [12]. Specifically, we adopt split conformal prediction [25], which involves designing a conformity score to measure the reliability of predictions, collecting a calibration set to establish standards for comparing conformity scores during testing, and determining whether to include a prediction result in the final prediction set. In the context of deep learning, the conformity score is defined as the probability output of models as this reflects its confidence in the answers it produces.

In our framework, we adopt p_k as conformity scores, and only Option_k with p_k higher than a computed threshold are included in the calibrated prediction set. To handle different probability distributions for **Options A** and **Options B**, we define two calibration sets: $\mathcal{Z}_{\text{cal}}^{A} = \{z_k = (\text{'A'}, p_k) \mid k = 1, ..., N_k\}$ and $\mathcal{Z}_{\text{cal}}^{B} = \{z_k = (\text{'B'}, p_k) \mid k = 1, ..., N_k\}$. To collect the calibration sets, we sample (observed_object, target_object) pairs from 20 diverse HM3D scenarios and generate ground truth labels. To handle scenarios where multiple objects may be relevant, we formally define the ground truth for calibration as the single option with the highest LLM confidence score, following the procedure in [12]. For each labeled pair, we apply the LLM reasoning process described in Section V.A to obtain probability values p_A and p_B, which form the calibration sets $\mathcal{Z}_{\text{cal}}^{A}$ and $\mathcal{Z}_{\text{cal}}^{B}$, respectively. These calibration datasets represent samples from the underlying distributions $\mathcal{D}_{\text{cal}}^{A}$ and $\mathcal{D}_{\text{cal}}^{B}$ for spatial co-occurrence judgments in household environments. Let $\mathcal{D}_{\text{test}}^{A}$ and $\mathcal{D}_{\text{test}}^{B}$ represent the unknown distributions of option-probability pairs for a new scenario, which are assumed to be exchangeable with $\mathcal{D}_{\text{cal}}^{A}$ and $\mathcal{D}_{\text{cal}}^{B}$.

In our MM-EQA setting, we argue that calibration and test samples can be treated as independent and identically distributed (i.i.d). Let (s_i, p_i) represent the i-th sample, where s_i denotes the spatial relationship (A or B) and p_i denotes the LLM's probability output. Our i.i.d. assumption states:

$$P((s_1, p_1), (s_2, p_2), \ldots, (s_n, p_n)) = \prod_{i=1}^{n} P(s_i, p_i), \quad (1)$$

where $(s_i, p_i) \sim F$ for all i, and F is a fixed joint distribution over spatial relationships and probability values. This assumption holds because the calibration process satisfies both independence and identical distribution requirements: 1) Calibration scenarios are randomly sampled from the HM3D household dataset without temporal or spatial dependencies; 2) Each (observed_object, target_object) evaluation represents an independent semantic judgment; 3) LLM's probability outputs for different object pairs are not causally related; 4) Spatial co-occurrence patterns represent typical household organization principles; 5) The LLM's semantic understanding remains consistent across scenes; 6) The underlying distribution F of object spatial relationships is preserved across the dataset; and 7) During testing, each robot generates semantic values for target and relevant objects based solely on its current view, without considering past frames, which ensures spatial and temporal independence required for conformal prediction. Fulfilling these conditions ensures that the prerequisite for applying conformal prediction is met. While CP typically requires only exchangeability of data, our i.i.d. assumption provides stronger theoretical guarantees. Approximately $1/\delta$ calibration samples are needed for reliable coverage [26], and 20 scenarios provide adequate data to achieve typical confidence levels (e.g., $\delta = 0.05$).

For a new test sample $z_{\text{test}} = (\text{'A' or 'B'}, p_{\text{test}})$, we use CP to construct a prediction set $\mathcal{C}(z_{\text{test}})$ using a calibrated score threshold, p_{thres}. This threshold is set to the $1-\epsilon_1$ quantile of the scores in the calibration set, where ϵ_1 is a user-defined miscoverage rate (e.g., $\epsilon_1 = 0.05$ for 95% confidence) that influences the size of the prediction set $\mathcal{C}(\cdot)$. The prediction set is then formed by including all options whose scores, p_{test}, meet or exceed this threshold. This construction provides the

formal marginal guarantee that the true label is contained within the prediction set with high probability:

$$P(z_{\text{test}} \in \mathcal{C}(z_{\text{test}})) \geq p_{\text{thres}} \tag{2}$$

A higher p_{thres} leads to more confident options being shared during communications, and vice versa. The corresponding $O^i_{\text{relevant},j}$ and $O^i_{\text{target},j}$ of the correct options along with approximate positions of correct options are used to generate the message ζ^i using the template: "I see {relevant object} that may be relevant to your target {true target}, and {possible target object} may be your target at {position}." If neither object is identified, the robot will not send a message.

C. Exploration with Communication

To guide exploration, we first compute the semantic values (SV) $\text{SV}^i_{\text{no-com},j}$ for a 2D grid map as per [2], ignoring the messages received from other agents. We denote P as a set of frontier points identified by the VLM at the current pose—locations on the boundary of the explored and unexplored regions [27]. The SV of $\hat{p} \in P$ is denoted as $\text{SV}^i_{\hat{p},j}$. Upon receiving a message $\zeta^{\hat{i}}$, the SV for each $\hat{p} \in P$ based on communication is adjusted based on the counts of the relevant and target objects, scaled by temperature τ_1 and τ_2:

$$\text{SV}^i_{\text{com},\hat{p},j} = \log\left(\tau_1 \text{Num}(O^i_{\text{relevant},j}) + \tau_2 \text{Num}(O^i_{\text{target},j})\right), \tag{3}$$

where τ_1 and τ_2 weight the counts of relevant and target objects, respectively, balancing indirect semantic cues and direct task information in the communicated semantic value. For each task, the SV is defined as:

$$\text{SV}^i_{\hat{p},j} = \max(\text{SV}^i_{\text{no-com},\hat{p},j}, \text{SV}^i_{\text{com},\hat{p},j}) \tag{4}$$

The semantic value at position $\hat{p}$ is set to the maximum of the agent's own estimate and the value from received messages, favoring the most informative source. The average of all SV^i_j values is computed to determine the final semantic value:

$$\text{SV}^i_{\text{final},\hat{p}} = \frac{1}{N_q} \sum_{j=1}^{N_q} \text{SV}^i_{\hat{p},j} \tag{5}$$

If $\text{SV}^i_{\text{final},\hat{p}} = 0$, the robot randomly selects one of the frontier points to continue exploration. After moving, the robot applies Volumetric Truncated Signed Distance Function (TSDF) Fusion [28], [29] to update the occupancy of the voxels in M, which marks them as explored or unexplored based on depth images I^i_d. This data is projected onto a 2D grid map matching the size of SV maps. The robot then uses Frontier-Based Exploration (FBE) [27] for path planning, where higher SV points indicate the surrounding areas likely to be more informative. To improve exploration efficiency, Gaussian smoothing spreads each SV point's value across its surrounding region, enabling smoother navigation and helping robots prioritize areas with high information value.

D. Confidence Check

The robot performs a confidence check to determine if it has sufficient certainty to answer an embodied question. Here, we prompt the VLM to generate answer prediction probabilities over the four possible answers from {'A', 'B', 'C', 'D'}, along with a question-image relevance score. This relevance score reflects the likelihood that the robot's current view contains the information needed to answer the question. The VLM's probability of predicting "Yes" to the prompt "Consider the question. Are you confident about answering the question given the current view?" is regarded as the question-image relevance score Rel^i_j for question q^i_j. The set of answer probabilities is defined as $\{\text{Ans}^i_j(L) \mid \ L \in \{\text{'A', 'B', 'C', 'D'}\}\}$. Both Rel^i_j and $\text{Ans}^i_j(L)$ are bounded within the interval $[0, 1]$. The answer is considered confident, and subsequent exploration will exclude question q^i_j, if the following condition is met: $\exists! L \in \{\text{'A', 'B', 'C', 'D'}\}$ s.t. $\text{Ans}^i_j(L) \times \text{Rel}^i_j > 1 - \epsilon_2$, where ϵ_2 is a user-defined threshold that reflects the confidence level needed for termination. This threshold ϵ_2 is applied to all robots when determining if an answer can be accepted.

E. Stop Criterion

In our setup, the robot terminates its exploration once it has answered all the questions assigned to it, either through its own analysis based on its observations and reasoning or with answers provided by the other robots. Alternatively, exploration stops when the maximum allowed time is reached.

V. Experiments

A. MM-EQA Benchmark

We create an MM-EQA benchmark based on the HM3D dataset [13] to evaluate our framework and baseline methods, which provides photo-realistic, diverse 3D indoor scenarios. The corresponding questions are challenging and require semantic reasoning. Different from the existing EQA datasets with a single question per scene, we generate six questions for each scene and their corresponding answers to support the multi-agent multi-task setting. The embodied questions can be categorized into the following five types:

1) **Location:** This type asks about the location of an object, e.g., "Where have I left the cushion? A) At the corner of the bedroom B) In the hallway C) Near the basketboard D) Next to TV in the living room".
2) **Identification:** This type asks about identifying the property of an object, e.g., "What bath mat is in the bathroom? A) Red B) White C) Black D) Gray".
3) **Counting:** This type asks about the number of objects, e.g., "Did I leave any cues or balls on the pool table? A) None B) One C) Two D) Three".
4) **Existence:** This type asks if an object is present at a certain location, e.g., "Have I put utensils and napkins on the dining table? A) Yes B) No".
5) **State:** This type asks about the state of an object, e.g., "Is the washing machine turned on? A) Yes B) No".

We use GPT-4V to generate the task questions and corresponding target objects, with human oversight refining them. Our benchmark includes 420 embodied question tasks across 70 scenarios for two robots, each assigned three tasks for testing. We allocate an additional 20 scenarios to create the calibration dataset for conformal prediction. We use Habitat [30] as the base simulator, which loads 3D scenes.

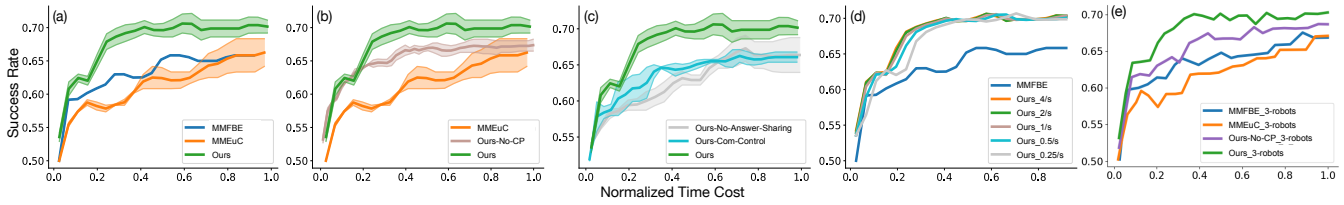

Fig. 3. The diagrams of SC vs. NTC on our MM-EQA dataset. (a)–(d) show the results for a 2-robot team. (a) The comparison between our method and baselines. (b) The ablative comparison of communication and conformal prediction modules. (c) The ablative comparison of the number of objects in the communication messages and the answering sharing mechanism. (d) The ablative comparison of baselines with our method at message-sending speeds of 0.25, 0.5, 1, 2, and 4 messages per second. (e) A scalability analysis comparing our method and baselines using a 3-robot team.

B. Evaluation Metrics and Baselines

We evaluate the performance of our method using two metrics: **Success Rate (SR)**, which measures the proportion of correct answers across all questions assigned to the robots, and **Normalized Time Cost (NTC)**, which quantifies the normalized time taken from the start of navigation to the completion of tasks by all the robots. The time cost consists of two components: robot movement and message sending. To analyze the impact of communication latency, we experiment with several different message-sending speeds. By default, we set the robot's movement speed to $1\,\mathrm{m/s}$ and the message-sending speed to one per second to maintain consistency and comparability across experiments.

We compare our method against the following baselines and ablation settings:

1) **MMFBE:** A multi-agent multi-task frontier-based exploration method extended from the canonical FBE [27]. MMFBE employs a VLM to answer questions but does not use it for semantic mapping during exploration and does not involve communication.
2) **MMEuC:** A multi-robot multi-task extension of the Explore Until Confident (EUC) framework [2], where robots operate independently without communication.
3) **Ours-No-CP:** An ablation that allows communication but omits conformal prediction, enabling evaluation of its contribution to calibrating LLM confidence.
4) **Ours-Com-Control:** An ablation that controls the number of objects included in communication messages. To match the effect of CP, we fix the number of objects to the same level as with CP and randomly sample from observed objects.
5) **Ours-No-Answer-Sharing:** An ablation where robots exchange observations and calibrated predictions relevant to each other's tasks but are not sharing answers, isolating the impact of answer-sharing on performance.

C. Implementation Details

We use the same VLM (Prismatic-VLM-13B) [31] as in [2]. Since our setting requires probability outputs from the LLM, we cannot use state-of-the-art closed-source models like GPT-4V. Instead, we employ LLaMA3-8B-instruct, a smaller open-source model. We set the temperature of LLM to 0.7. We set the τ_1, τ_2 to be 1.0 and 10.0, respectively. The non-conformity threshold p_{thres} is computed as the $(1-\epsilon_1)$ quantile of the calibration set probabilities, where ϵ_1 controls the prediction set size. In our setup, this corresponds to a 0.6 quantile for Option A and a 0.82 quantile for Option B. We conduct all experiments using two NVIDIA 6000 Ada Generation GPUs, with VLM and LLM on separate devices.

D. Results and Analysis

Communication Effectiveness. We demonstrate the effectiveness of our communication module through the average success rate achieved when the robots are allowed to explore within different time horizons, which is shown in Fig. 3(a). Our method significantly outperforms the baseline, MMFBE, by achieving an SR of 0.68 at an NTC of 0.4, compared to MMFBE's SR of 0.65 at an NTC of 0.8, effectively doubling the efficiency. This substantial efficiency gain comes from our method's ability to share relevant, calibrated information, allowing agents to prioritize the exploration of important areas. In contrast, the "MMEuC" baseline underperforms even MMFBE because it lacks communication between agents. Without communication, robots are unable to coordinate their efforts or share insights, leading to redundant or inefficient exploration paths. Additionally, "Ours" achieves the highest final success rate at convergence, which implies the effectiveness of our communication strategy. In terms of actual time spent, "Ours" completes all questions in an average of 445 seconds, compared to 594 seconds for MMFBE, demonstrating faster task completion.

The superior performance of our approach is attributed to the strategic exchange of information, which enables agents to collectively explore relevant objects rather than randomly searching without context. This is a clear advantage over MMFBE which lacks heuristic-based guidance and operates without the benefit of collaborative data sharing. It is also notable that as the NTC increases and more objects are explored, there is greater diversity in the LLM's output, consequently causing an increase in the standard deviation of the results. There is no deviation in MMFBE since FBE is a deterministic rule-based method.

In Fig. 3(c), the comparison with the "Ours-No-Answer-Sharing" baseline shows the impact of sharing answers to other robots' questions. Without this ability, a robot spends additional time exploring answers to questions that could be addressed by its partners. This lack of efficiency is reflected in the higher NTC and a lower SC compared to the full "Ours" configuration. Consequently, the absence of this capability results in slower convergence and reduced overall task efficiency, as each agent must independently answer questions without fully leveraging high-confidence information shared by its partners. By enabling answer sharing, our method ensures that high-confidence shared information is

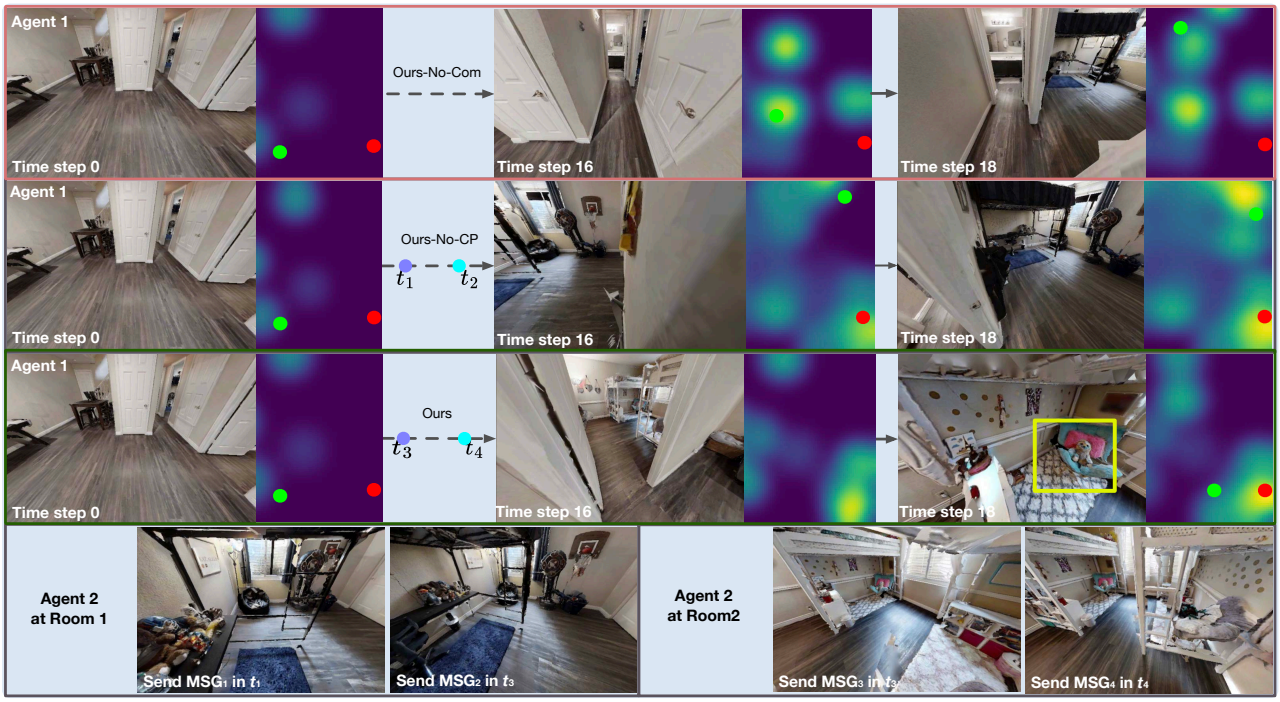

Fig. 4. The comparisons of robot views and global SV maps among three methods. The red points represent the location of the target and the green points represent the position of the robot. Agents in the three methods start from the same pose. The question for this scenario is "Where is the red bear cushion?" For "Ours-No-CP" and "Ours", Robot2 separately explores the same rooms at different times and sends messages to Robot1. The detailed messages are as follows: MSG_1: I see a basketboard, dolls, black chair that may be relevant to your target red bear cushion, and dolls may be your target at $\{position1\}$. MSG_2: I see dolls that may be relevant to your target red bear cushion at $\{position2\}$. MSG_3: I see bed, red pillow on blue chair that may be relevant to your target red bear cushion, and red pillow on blue chair may be your target at $\{position3\}$. MSG_4: I see red pillow on blue chair that may be relevant to your target red bear cushion, and a red pillow on blue chair may be your target at $\{position4\}$.

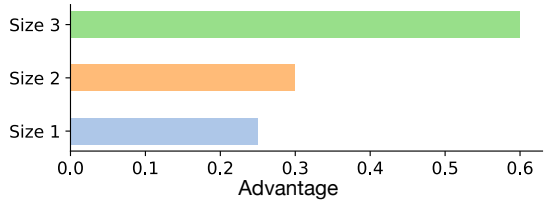

Fig. 5. The comparison of performance improvement in the environments with different sizes. The "Advantage" represents the difference between the NTC of "Ours" and the NTC of MMFBE, calculated as Advantage $=$ $\text{NTC}_{\text{Ours}} - \text{NTC}_{\text{MMFBE}}$. Size 1 represents scene area $L \times W < 150\,\text{m}^2$. Size 2 represents $150 \leq L \times W < 250\,\text{m}^2$. Size 3 represents $L \times W \geq 250\,\text{m}^2$.

leveraged to facilitate task completion.

Confidence Calibration with Conformal Prediction. We illustrate the importance and effectiveness of CP in Fig. 3(b). Without CP, the "Ours-No-CP" baseline achieves performance similar to "MMEuC", showing that uncalibrated outputs mislead agents with ineffective or erroneous information, leading to exploration in less relevant areas and reduced efficiency. In contrast, the calibrated communication in "Ours" ensures that messages are reliable, improving task success. Fig. 3(c) shows that the "Ours-Com-Control" configuration, which sends more frequent but less relevant information, fails to improve efficiency, implying that the *quality* of shared information is more critical than its *quantity*. Ineffective communication, providing more information with less relevance in "Ours-Com-Control", results in worse performance and slower convergence, with the final SR similar to those in the "MMEuC" configuration.

We demonstrate a representative test case in Fig. 4. In the "MMEuC" configuration, the SV map modes diffuse slowly and do not cover the important areas, which leads to a low probability of the robots finding target objects. This is because they tend to move to irrelevant areas rather than exploring the correct ones, decreasing the chances of finding targets. Although the modes of the SV map in "Ours-No-CP" diffuse more rapidly than in "Ours", the uncalibrated communication causes the robot to navigate to the wrong room due to misleading information from miscalibrated messages. In contrast, "Ours" effectively updates the ego robot's semantic value map, guiding the robot efficiently and demonstrating the importance of calibrated communication.

Impact of Scene Size. Fig. 5 further highlights how scene size affects the performance of "Ours" compared to the MMFBE baselines. As the scene area increases, the advantage of our communication method becomes more pronounced. In larger scenes (Size 3), our method achieves an average NTC improvement of 0.6 over MMFBE, showcasing a substantial efficiency gain due to enhanced information sharing and coordinated exploration. The MMFBE baseline struggles more as scene size increases because it relies on rule-based, non-communicative exploration, which lacks adaptability to larger, more complex environments. Our method, leveraging calibrated and relevant shared information, consistently enhances exploration efficiency and task success, which demonstrates its robustness and scalability.

Impact of Communication Latency. We evaluate task success under different message-sending speeds to analyze the effect of communication latency. Fig. 3(d) shows that

higher sending speeds lead to faster increases in success rates in early stages by enabling faster information exchange. After sufficient exploration, the final success rates under different speeds become almost identical. Notably, our approach outperforms the MMFBE baseline across all message-sending speeds, demonstrating its effectiveness regardless of latency.

Scalability Analysis. We evaluate scalability by extending to a three-robot team. As shown in Fig. 3(e), "Ours" achieves a faster increase in SR with respect to NTC compared to the baselines. Conversely, "Ours-No-CP" suffers a decline in early-stage SR due to the increased presence of irrelevant information. While all methods benefit from more agents over time, our method consistently completes tasks more efficiently and scales with minimal computational overhead.

VI. Conclusion

To improve the communication efficiency of decentralized LLM-based multi-agent systems, we propose an effective natural language communication framework with LLMs. Due to the hallucinations of LLMs and their potentially negative effects on overall efficiency, our framework incorporates conformal prediction to calibrate the confidence of outputs, which filters out overconfident LLM-generated outputs and makes the information in communication effective. The experimental results demonstrate that our method significantly improves exploration efficiency and the overall success rate, particularly in larger scenes. In this work, we develop our method in both two- and three-agent cooperative settings through efficient communication. Future work will involve scaling our method to more complex deployments with significantly larger robot teams and environments.

References

[1] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1–10.

[2] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until confident: Efficient exploration for embodied question answering," *Robotics: Science and Systems*, 2024.

[3] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud *et al.*, "Openeqa: Embodied question answering in the era of foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 488–16 498.

[4] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6309–6318.

[5] J. Yao, X. Zhang, Y. Xia, A. K. Roy-Chowdhury, and J. Li, "Towards generalizable safety in crowd navigation via conformal uncertainty handling," in *Conference on Robot Learning (CoRL)*, 2025.

[6] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[7] Y. Zhang, S. Yang, C. Bai, F. Wu, X. Li, X. Li, and Z. Wang, "Towards efficient llm grounding for embodied multi-agent collaboration," *arXiv preprint arXiv:2405.14314*, 2024.

[8] S. Rasal, "Llm harmony: Multi-agent communication for problem solving," *arXiv preprint arXiv:2401.01312*, 2024.

[9] S. J. Mielke, A. Szlam, E. Dinan, and Y.-L. Boureau, "Reducing conversational agents' overconfidence through linguistic calibration," *Transactions of the Association for Computational Linguistics*, 2022.

[10] G. Hao, J. Wu, Q. Pan, and R. Morello, "Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks," *Scientific Reports*, vol. 14, no. 1, p. 16375, 2024.

[11] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[12] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," in *7th Annual Conference on Robot Learning*, 2023.

[13] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[14] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better llm-based evaluators through multi-agent debate," in *The Twelfth International Conference on Learning Representations*, 2024.

[15] X. Zhang, H. Qin, F. Wang, Y. Dong, and J. Li, "Lamma-p: Generalizable multi-agent long-horizon task allocation and planning with lm-driven pddl planner," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[16] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin *et al.*, "Metagpt: Meta programming for a multi-agent collaborative framework," in *The Twelfth International Conference on Learning Representations*, 2024.

[17] J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" National Bureau of Economic Research, Tech. Rep., 2023.

[18] W. Hua, L. Fan, L. Li, K. Mei, J. Ji, Y. Ge, L. Hemphill, and Y. Zhang, "War and peace (waragent): Large language model-based multi-agent simulation of world wars," *arXiv preprint arXiv:2311.17227*, 2023.

[19] Y. Chen, J. Arkin, Y. Zhang, N. Roy, and C. Fan, "Scalable multi-robot collaboration with large language models: Centralized or decentralized systems?" in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4311–4317.

[20] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[21] J. Zhang, Y. Hou, R. Xie, W. Sun, J. McAuley, W. X. Zhao, L. Lin, and J.-R. Wen, "Agentcf: Collaborative learning with autonomous language agents for recommender systems," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 3679–3689.

[22] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig *et al.*, "Sotopia: Interactive evaluation for social intelligence in language agents," in *The Twelfth International Conference on Learning Representations*, 2024.

[23] J. Wang, G. He, and Y. Kantaros, "Safe task planning for language-instructed multi-robot systems using conformal prediction," *arXiv preprint arXiv:2402.15368*, 2024.

[24] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[25] M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: a unified review of theory and new challenges," *Bernoulli*, vol. 29, 2023.

[26] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, and M. Pavone, "Sample-efficient safety assurances using conformal prediction," *The International Journal of Robotics Research*, 2024.

[27] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997.

[28] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136.

[29] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1802–1811.

[30] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. Clegg, M. Hlavac, S. Y. Min *et al.*, "Habitat 3.0: A co-habitat for humans, avatars, and robots," in *The Twelfth International Conference on Learning Representations*, 2024.

[31] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*, 2024.