2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SiSCo: Signal Synthesis for Effective Human-Robot Communication Via Large Language Models

Shubham Sonawani, Fabian Weigend and Heni Ben Amor

***Abstract*— Effective human-robot collaboration hinges on robust communication channels, with visual signaling playing a pivotal role due to its intuitive appeal. Yet, the creation of visually intuitive cues often demands extensive resources and specialized knowledge. The emergence of Large Language Models (LLMs) offers promising avenues for enhancing human-robot interactions and revolutionizing the way we generate context-aware visual cues. To this end, we introduce SiSCo– a novel framework that combines the computational power of LLMs with mixed-reality technologies to streamline the creation of visual cues for human-robot collaboration. Our results show that SiSCo improves the efficiency of communication in human-robot teaming tasks, reducing task completion time by approximately 73% and increasing task success rates by 18% compared to baseline natural language signals. Additionally, SiSCo reduces cognitive load for participants by 46%, as measured by the NASA-TLX subscale, and receives above-average user ratings for on-the-fly signals generated for unseen objects. To encourage further development and broader community engagement, we provide full access to SiSCo's implementation and related materials on our GitHub repository.**[1]

I. Introduction

Human-robot collaboration (HRC) relies on clear and intuitive *communication* channels between the human participants and their robotic counterparts. This foundational clarity is essential for both entities to understand each other's intentions and, in turn, engage in safe and effective interactions. Building upon this insight, a diverse set of communication modalities has been explored in the field of HRC, including text, auditory cues, gestures, and visual signals [1]. Among these modalities, visual signals stand out due to their ability to (a) quickly capture attention and (b) instantly transmit a multitude of information. Accordingly, in recent years several approaches have been proposed that leverage visual cues within a mixed-reality environment to enhance HRC [2]–[5] The proliferation of affordable output devices for virtual and augmented reality (e.g., Meta Quest, Microsoft HoloLens, Apple Vision Pro) has further increased the interest in visual forms of communication for human-robot teaming.

While visual signals excel in quickly conveying a variety of information, their design and production are neither straightforward nor effortless and usually require human experts. The field of human-centered computing encompasses a dedicated area of study known as *Visual Languages* [6], which focuses on principles for creating impactful visual information. In general, the process of creating these visual cues requires careful consideration in order to balance the need for clarity and intuitiveness against the resources and expertise needed for their development. Previous approaches for visual signaling in robots resorted to a predefined grammar [7], i.e., a context-free language, or special-purpose authoring tools for the production of visual cues [8]. While intuitive, these approaches do not allow for the extemporaneous generation of novel visual signals.

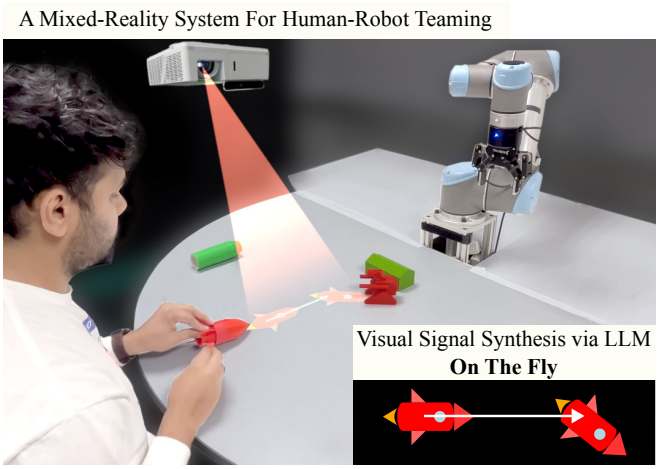

Fig. 1: A participant engaging in a human-robot teaming task, with the SiSCo framework mediating by delivering visual signals via a mixed reality interface.

On the other hand, recent work on Large Language Models (LLMs) shows the impressive ability in communicating with a human in a free-form fashion, i.e., without the need for a strict grammar or template. Particularly notable is the capacity of LLMs to comprehend the nuances of context in prompts provided by users and to generate coherent, contextually relevant textual responses [9], [10] which showcases LLMs reasoning capabilities. LLMs draw their capabilities from training on expansive textual datasets, comprising a diverse array of internet sources, including web pages, code repositories, and scholarly publications. By assimilating this wide spectrum of human-expressed knowledge, the models effectively encapsulate vast informational breadth into their trained weights.

In this paper, we address the question of whether it is possible to bridge the divide between textual and visual communication for the purposes of efficient human-robot teaming. Given a large language model, is there way we can leverage the embedded human knowledge to expertly generate signals for human robot collaboration? This paper specifically investigates the potential of LLMs to synthesize both natural language and visual signals that can adapt

S. Sonawani, F. Weigend, and H. Ben Amor are with the School of Computing and Augmented Intelligence, Arizona State University {sdsonawa, fweigend, hbenamor}@asu.edu

[1]https://github.com/ir-lab/SiSCo

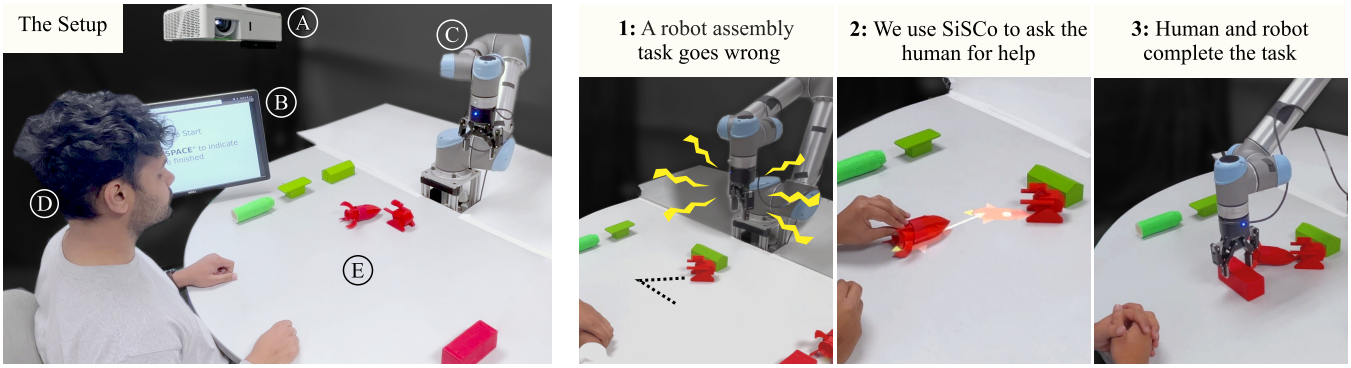

Fig. 2: Left: The physical setup of the teaming task: The robot places objects on the tabletop surface environment. When the robot needs help, it uses SiSCo to present synthesized signals through a projector (A) or a monitor (B) to the human. **Right:** The task procedure during the human-robot teaming task.

to changes in the environment, thereby enabling them to effectively convey robot intentions to users. An illustrative example of SiSCo deployed in a human-robot teaming scenario is presented in Figure 1.

To this point, the contributions made by this paper are outlined as follows:

- The development of **SiSCo** (**Si**gnal **S**ynthesis for Effective Human-Robot **Co**mmunication), a novel framework integrating Large Language Models with mixed reality to produce legible visual signals on the fly, enhancing human-robot collaborative tasks.
- An empirical study involving human participants to rigorously assess the impacts of **SiSCo** on the enhancement of communication in teaming task performance, the alleviation of cognitive load, and the demonstration of its strong generalization capabilities in varied user inputs.
- Provision of the open-source repository for the **SiSCo** framework, including full system implementation details, to encourage widespread adoption and allow for reproducibility.

II. Related Work

Effective communication plays a crucial role in human-robot interaction (HRI), with signaling methods serving as a critical component in clarifying intentions and reducing ambiguities [1], [11]. Robots communicate with their human counterparts through a variety of signals, predominantly motion cues, gestures, and sounds. The clarity or "legibility" [12] of these signals is essential to facilitate understanding of robot behavior and intentions.

Several studies have shown that both implicit and explicit signaling methods can significantly improve the communication gap in HRI tasks [12], [13]. For example, implicit cues such as haptic feedback or eye movements can be particularly useful in close-proximity physical tasks [14]–[16]. Implicit cues can be communicated alongside the original robot behavior but may require an adaptation phase, i.e., the human partner may have to learn the meaning of a given cue and how to best respond to it. By contrast, explicit cues provided via vision or language may require an additional communication step but are often easier to interpret for a human partner, e.g., the robot interrupting the task to provide a distress signal via language or through a picture [17]. Immersive technologies, including virtual and mixed reality, have been adopted to deliver high-fidelity visual information and improve collaboration between humans and robots [18]. While virtual reality has the potential to increase the effectiveness and efficiency of collaborative tasks, it may cause user fatigue or nausea [19]. By contrast, mixed-reality techniques like projecting visual signals onto the real-world workspace can facilitate a simpler and more efficient communication mode in collaborative settings [2]. All such approaches, however, require consistent and clear visual signals to achieve the intended effect [20]–[22]. The most common way to create visual signals for a new scenario is to design them in a manual process. In turn, they can be reused as long as the task domain remains same. Alternatively, the work by [7] proposed the concept of a domain-specific visual language, which utilizes a set of composable visual signals. Central to this proposition was the introduction of a robust grammar that served as an underlying generation mechanism. While the grammar itself stands as a well-conceived framework for consistent interpretation of signals, its application imposes an inherent rigidity. Specifically, visual signals, as defined by the established grammar, remain static and cannot be adjusted as the task and the environment evolves.

This stands in stark contrast to the flexibility of LLMs which have recently led to a revolution in human-centered computing [23], [24]. LLMs provide a non-rigid communication interface, that allows humans to interact with a large corpus of textual data. Recent work, has shown that such textual interaction can also be leveraged to engage with robot partners [25], [26]. Most importantly, LLMs have demonstrated an impressive adaptability to perform various tasks in a few-shot or even zero-shot manner. Accordingly, they are able to adapt to new tasks with minimal or no human effort. In the remainder of this paper, we investigate how the flexibility of LLMs can be used to generate visual signals for robotics tasks in a systematic and reliable manner.

III. Methodology

We present **Si**gnal **S**ynthesis for Effective Human-Robot **Co**mmunication (SiSCo) – a system that utilizes LLMs

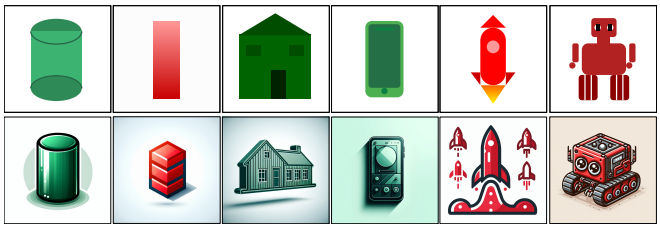

Fig. 3: Object signals generated from SiSCo (Top Row) and Dalle-3 (Bottom Row) for same input prompt

to understand the task context and synthesize meaningful visual signals. Contextual information about the state of the collaboration task such as environmental descriptors (e.g., table dimensions) and problem statements (e.g., potential assembly errors by the robot) serve as input to the system. In turn, a visual representation is synthesized and projected into the scene to affect human behavior. While the projection can be performed via any display device (e.g., head-up display), we focus in the remainder of this paper on a projection-based methodology. Figure 2 depicts the overall physical setup of our system.

Central to our approach is the generation of Scalable Vector Graphics (SVG) code, which is proficiently produced by the LLM framework for the creation of visual signals. The use of LLM is strategic, given its adeptness at contextual understanding and its ability to provide only the most pertinent information. While Vision Foundation Models [27] offer comprehensive visual representations of objects and scenes, they deliver an abundance of data in the form of high-dimensional RGB information, which may surpass the necessity for recognizing basic objects like a toy humanoid robot on a work surface (Figure 3). In stark contrast, SVG not only provides an efficient abstraction of these high-resolution images through simple geometric shapes and colors but also represents a human-interpretable file format. In addition, SVG files have the added advantage of being infinitely scalable without loss of quality; they can be rasterized to match the full dimensional range of traditional RGB arrays as required, ensuring compatibility with diverse display resolutions and optimizing system resource utilization for real-time applications in human-robot interaction.

A. System Overview

A key component of our system is a mixed-reality setup, known as intention projection [28]–[30], to communicate synthesized visual signals to the human. The setup employs a projector to superimpose visual information onto the real-world setting for clear communication of intentions and instructions. Figure 2 explains overall setup where the user, labeled as D, cooperates with the Universal Robot (UR5), indicated as C, to complete an assembly task on a table, designated as E. The objects required for the assembly are positioned on both sides of the table. If the robot experiences a malfunction, it initiates a query to the SiSCo framework. SiSCo then processes this query and conveys the synthesized visual signals to the user. These signals can be displayed using two methods: a projector (A) targeting the tabletop

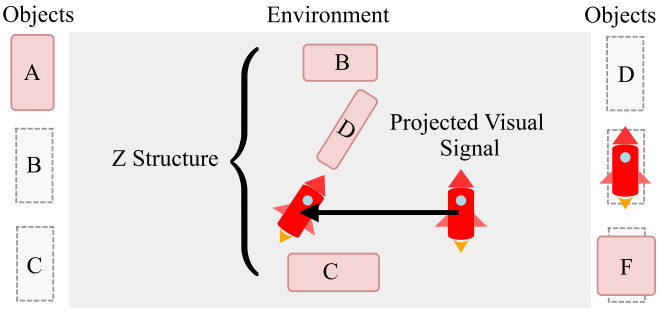

Fig. 4: The teaming task as a schematic: The robot assembles a Z-shaped structure on the tabletop using the objects B, C, D and the red rocket. The robot placed the first three objects but malfunctioned when placing the red rocket. SiSCo projects a visual signal to instruct the human to assist.

surface, enhancing spatial relevance, or an adjacent monitor (B) for a natural or visual signals.

The teaming task in this environment require cooperation between a robot, acting as the assembler, and a human, serving as an observer and assistant. In the example procedure depicted on the right in Figure 2, the goal is to arrange the objects in a Z-shaped structure. In Figure 2 (Step-1), the robot has already placed two objects, but intentionally (simulated) failed to pick up the rocket. SiSCo then synthesizes a visual signal and projects it onto the table to ask the human for help (Step 2). The human places the rocket, which enables the robot to complete the Z-shaped structure (Step 3).

B. Definitions

This section aims to formalize the components related to the teaming task, including the workspace, interacting objects, and structures. Below we provide definitions for each of these components:

Environment: We define the operational environment as a 1.4m by 0.7m rectangular workspace on a tabletop, further subdivided into a 1400 $\times$ 700 cell grid for visual signals provided by projection based Mixed Reality (MR), as illustrated in Figure 4. This grid, serving as a digital canvas, enables precise MR-based visual cues for robot and human interaction within this space. At the center of this digitally augmented environment, the robot is tasked with assembling structures from various objects, simulating complex interaction scenarios. This setup optimizes the use of space for human-robot collaboration and tests the robot's ability to interpret and act upon MR signals in real-time.

Objects: We position the objects at the left and right edges of the environment. Each physical object has a maximum width of 5 cm and a length of 13 cm. Their height varies between 6 cm and 9 cm. Additionally, each object has been assigned a description, denoted by Δ, and a color, represented by Θ.

Structure: One by one the robot picks and places objects to assemble alphabet-inspired structures in the center of the environment. In the example in Figure 4, the structure has a Z-shape.

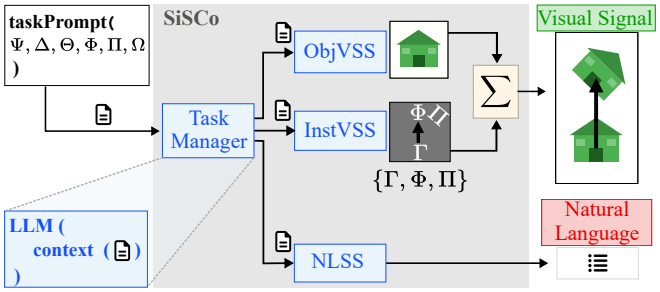

Fig. 5: The architecture of our Signal Synthesizing Communication System (SiSCo). It takes in a task prompt and produces visual and natural language signals.

```
<svg width="250" height="250" xmlns="http://www.w3.org/2000/svg">
  <!-- White background -->
  <rect width="100%" height="100%" fill="white" />
  <!-- Robot arm base -->
  <rect x="100" y="200" width="50" height="20" fill="#c0c0c0" />
  <!-- Robot arm first segment -->
  <rect x="120" y="130" width="10" height="70" fill="#4682b4" />
  <!-- Robot arm second segment -->
  <rect x="120" y="80" width="10" height="50" fill="#4682b4" />
  <!-- Robot arm joint circles -->
  <circle cx="125" cy="130" r="5" fill="#c0c0c0" />
  <circle cx="125" cy="80" r="5" fill="#c0c0c0" />
  <!-- Robotiq two-finger gripper -->
  <rect x="115" y="60" width="20" height="20" fill="#c0c0c0" />
  <rect x="110" y="40" width="5" height="20" fill="#4682b4" />
  <rect x="135" y="40" width="5" height="20" fill="#4682b4" />
  <!-- Gripper fingers -->
  <rect x="110" y="30" width="5" height="10" fill="#c0c0c0" />
  <rect x="135" y="30" width="5" height="10" fill="#c0c0c0" />
  <!-- Gripper finger tips -->
  <circle cx="112.5" cy="30" r="2.5" fill="#c0c0c0" />
  <circle cx="137.5" cy="30" r="2.5" fill="#c0c0c0" />
</svg>
```

Fig. 6: Raw SVG code output generated by the ObjVSS function.

Problem Formalization: While assembling the structure, the robot simulates having a problem and being unable to place an object at its intended goal position. We formalize the problem via six parameters. All parameter values are strings. The following example values define the Z-Problem in Figure 4, where the robot fails to place the rocket:

- Ψ : structure (e.g. "Z")
- Δ : object description (e.g. "Rocket")
- Θ : object color (e.g. "Red")
- Φ : goal position (e.g. "[496, 100]")
- Π : goal orientation (e.g. "35 deg")
- Ω : instruction: (e.g. "insert from right")

From these parameters, SiSCo then synthesizes a signal to make the human complete the task for the robot (Projected Visual Signal in Figure 4). As a result, signals are synthesized to guide the human on how to contribute to resolving a problem within the environment.

C. Signal Synthesis

To communicate task goals and robot intentions to the human, SiSCo utilizes three signal modalities: Natural Language Signals (NLS), Visual Signals on a Monitor (VSM), or Visual Signals via Intention Projection (VSIntPro).

- **Natural Language Signal (NLS)**: This mode leverages the LLM to generate textual instructions. The human participant is provided with a set of succinct on-screen directives that details task properties and the object to be manipulated.
- **Visual Signal on Monitor (VSM)**: In this mode, SiSCo synthesizes a visual depiction of the required human intervention and displays it on the monitor.
- **Visual Signal via Intention Projection (VSIntPro)**: This mode employs the mixed-reality setup to signify the forthcoming object and its intended direction of manipulation. The signals are projected directly onto the tabletop.

The SiSCo system synthesizes these signals through hierarchical LLM queries and post-processing. As depicted in Figure 5, we formalize the hierarchical processing pipeline as four LLM function calls. Namely, the function calls are: 1) A Task Manager to subdivide the incoming prompt for following function calls, 2) A Natural Language Signal Synthesizer (NLSS) to summarize the task prompt in four bullet points, 3) An Object Visual Signal Synthesizer (ObjVSS) to process a prompt into an object icon and 4) An Instruction Visual Signal Synthesizer (InstVSS) to process a prompt into a visual instruction. All four LLM functions wrap a context around their input prompt, i.e., they prepend and append contextual information to the prompt before they query the model. Formally, we denote the prefix as PRE and the postfix as POST. The entire prompts including prefix and postfix are provided in our github repository. We describe their functions in the following sections:

Task Manager: As depicted in Figure 5, the input task prompt first reaches the Task Manager function, which then generates three distinct prompts for subsequent LLM function calls. The Task Manager wraps the pretext PRE_{TM} and posttext POST_{TM} around the input prompt and queries the LLM. Then, it uses regular expressions to parse the LLM return into three new prompts to be passed to NLSS, ObjVSS, and InstVSS.

NLSS: The Natural Language Signal Synthesizer (NLSS) takes the refined task prompt provided by the Task Manager and envelops it with PRE_{NLSS} and $\text{POST}_{\text{NLSS}}$. The resulting query compels the LLM to summarize the task description into a succinct set of four bullet points to be presented to the human.

ObjVSS: The Object Visual Signal Synthesizer (ObjVSS) generates the visual representation of the object in the form of SVG code. It receives a descriptive input prompt detailing the object description (Δ) and color (Θ) from the TaskManager. Subsequently, it augments the prompt with the context prefix $\text{PRE}_{\text{ObjVSS}}$ and postfix $\text{POST}_{\text{ObjVSS}}$, which delineate the specifications for the SVG creation, such as setting its target dimensions (e.g., 210×210 pixels) and specifying the background color. Examples of ObjVSS outputs in the form of SVG code snippet is shown in Figure 6 and are visualized in Figure 7.

InstVSS: For the Instruction Visual Signal Synthesizer (InstVSS), the Task Manager provides a prompt that encapsulates both the goal position (Φ) and goal rotation (Π) of the object to be manipulated. The contextual information $\text{PRE}_{\text{InstVSS}}$ and $\text{POST}_{\text{InstVSS}}$ obligate the LLM to contrive 1) suitable start position (Γ) and goal position in pixel coordinates 2) the SVG code that visualizes the trajectory, and 3) the object's orientation expressed in degrees, indicating

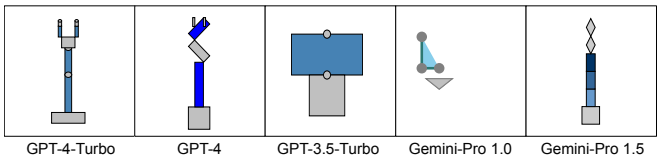

Fig. 7: Comparison of visual signals generated by different LLMs. ObjVSS input prompt: "*Generate an icon for an object with the description robotic arm with three degrees of freedom and parallel-jaw gripper and of blue and silver color.*"

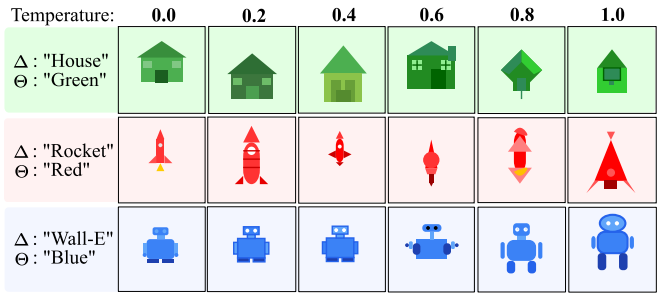

Fig. 8: Object representations can be synthesized reliably even with high-temperature values.

clockwise rotation from the vertical axis.

Sigma: The amalgamation of visual signals is conducted by the Sigma (Σ) function. Specifically, on an empty black canvas, the SVG output from the ObjectVSS is superimposed onto the designated start (Φ) and end goal (Γ) positions, as determined by the InstVSS. At the goal position, the SVG representation of the object is rotated by the specified angle Π. Subsequently, the trajectory SVG, also provided by the InstVSS, is superimposed to yield a comprehensive visual signal of the object's movement from the start to the goal position.

D. LLM Selection

Because SiSCo operates through text-based queries, it can interface with any LLM. We conducted initial experiments to select the most suitable and advanced LLM in the field for synthesizing reliable signals. Figure 7 depicts a representation of example outputs. It became evident from our initial comparison that, at the time of this work, OpenAI's GPT-4-Turbo exceeded the signal generation capabilities of the other. This compelled us to select GPT-4-Turbo as the signal synthesis component within the SiSCo framework for this work.

Acknowledging that GPT-4-Turbo demonstrates proficiency in the synthesis of signals, we further assessed its consistency in generating outputs across different *temperature* settings. For LLMs, the temperature parameter influences the stochastic nature of the output, as detailed in [31], [32]. A lower temperature setting results in more deterministic outputs, while a higher value facilitates greater randomness. The comparison depicted in Figure 8 illustrates the ObjVSS outputs for three distinct inputs, each defined by object descriptions (Δ) and color (Θ), across a temperature range from 0.0 to 1.0. Even at the highest temperature setting, the generated objects remain clearly representative of the input

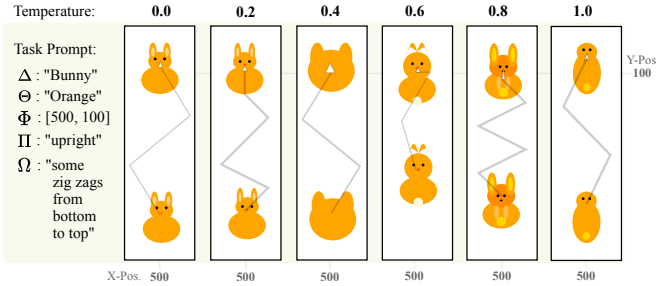

Fig. 9: Visual signal synthesis stays robust with increasing temperature. The goal position at [500, 100] is matched, icons resemble a bunny, and the trajectory shows a zig-zag pattern.

descriptions. This outcome underscores the robustness of the carefully designed LLM ObjVSS, demonstrating its ability to maintain consistency and accuracy under varying conditions.

Furthermore, we illustrate the robustness of the ObjVSS and InstVSS functions by examining the merged visual signals in Figure 9. Across various tested temperature settings, the target position (Φ) at [500, 100] is consistently attained, icons effectively depict an orange (Θ) bunny (Δ), and the trajectory displays an upward zig-zag pattern (Ω). Thus, we conclude that GPT-4-Turbo is well-suited for assessing the capabilities of signal synthesis for human-robot collaboration in our forthcoming experiments.

IV. Experiments

We designed an extensive human subject study to assess the quality of SiSCo for effective human robot communication. The study is subdivided into two parts: 1) A real-robot teaming task where a UR5 and a human assemble structures and 2) a visual signal evaluation questionnaire.

1) Human-Robot Teaming Task: As defined in Section III-A, the teaming task involved assembling structures at the center of the tabletop environment. During assembly, the robot got stuck and used SiSCo to ask the human for help. For this task, we used a set of six distinct objects: They included a red rocket, a red cuboid, a red Wall-E robot, a green mobile phone, a green house, and a green cylinder. Using these objects, participants encountered six assembly problems defined in Table I in randomized order and with random SiSCo signal modalities (NLS, VSM, and VSIntPro). We randomized problems such that participants completed two problems for each signal modality.

TABLE I: Test set definitions for our practical teaming task

Struct.	Object		Placement		
Ψ	Δ (Desc.)	Θ (Col.)	Φ (Pos.)	Π (Ori.)	Ω (Inst.)
S	Cuboid	Red	496, 262	90 deg	from bottom
Z	Rocket	Red	452, 306	45	from bottom
U	Wall-E Robot	Red	396, 336	same	from left
O	Cylinder	Green	598, 170	no change	from bottom
R	Mobile	Green	612, 414	-pi/4	insert from right
K	House	Green	496, 152	45 degrees	slide up from bottom

To evaluate the effect of SiSCo's signal modalities, we defined objective and subjective metrics. The objective metric one **OM1** captures failure and success rates, defining failure as: (1) manipulating the wrong object (Δ or Θ), (2) placing the object with an incorrect orientation (Π), (3) failure to

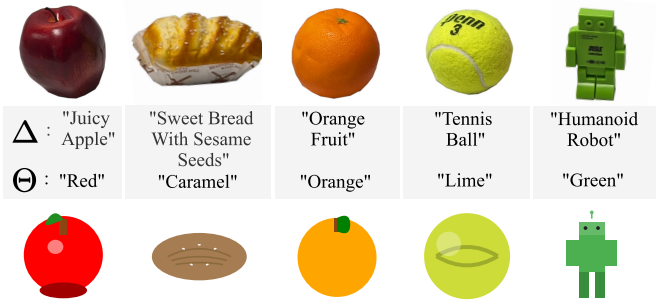

Fig. 10: Real objects, prompt object descriptions (Δ), prompt colors (Θ), and generated visual signals. These objects were used for the second part of the questionnaire study

follow the placing instruction (Ω), and (4) placement beyond 10 cm from the goal position (Φ). The second objective metric **OM2** quantifies task efficiency, measuring task completion time (duration from when the robot signaled being stuck to when the human placed the object, allowing the robot to continue) and comprehension time (duration from signal display to the participant touching an object). As the subjective metric **SM1** we utilized the NASA Task Load Index [33] to assess the participant's perceived mental demand and frustration when following SiSCo generated signals. We asked participants to provided their NASA Task Load Index scores after they encountered a signal type (NLS, VSM, or VSIntPro) for the first time. Further, we facilitated the System Usability Scale (SUS) [34] as the subjective metric **SM2** to assess the perceived overall effectiveness of SiSCo in the human-robot teaming task. Participants provided their SUS scores after completing all six assembly tasks.

2) Questionnaire: As the second part of our experiment, we asked the human to complete a questionnaire, which further evaluated SiSCo's signal efficacy, quality, and user preference. Our questionnaire had four parts. The first question asked participants to rank the effectiveness of each interaction mode (NLS, VSM, and VSIntPro) by assigning a score from 1 for the most effective to 3 for the least effective. We used the ranking as our subjective metric **SM3**.

The second questionnaire item assessed the object representation quality of ObjVSS outputs. We placed five real-world objects (depicted at the top in Figure 10) in front of the user. Then, we presented generated ObjVSS representations (bottom of Figure 10) in random order. Humans had to select the real object that corresponded to the signal and rate the signal representation on a scale ranging from -5 (unrecognizable) to 5 (ideal representation). We utilized selection success rate as the objective metric **OM3** and the rating as the subjective metric **SM4**. The third section of the questionnaire asked participants to input a string that corresponds to an image combining both ObjVSS and InstVSS elements. Specifically, the image required participants to provide an input representing the action, "*Insert a green-colored leek object from the left to the center of the image*." This section was designed to evaluate whether participants could accurately generate a signal prompt based solely on the visual representation, effectively reversing the typical process. For metric **SM5**, participants were then instructed to rate the re-engineered signal derived from their interpretation of the image.

In the final part of the questionnaire, participants were granted complete control over SiSCo and were asked to input prompts for any generic object, orientation and trajectory of their choice. After receiving the generated signal from SiSCo, participants provided the subjective metric **SM6** by rating the accuracy with which the signal represents the specified properties.

The underlying question behind the design of SiSCo is to evaluate its efficacy in improving communication between robot and human, and its ability to generate interpretable signals given novel inputs. To this point, we investigated the following hypotheses:

H1: Visual signals synthesized from SiSCo improve task performance compared to natural language signals.

H2: The cognitive load is lower in visual signal based modes compared to the natural language signal mode.

H3: SiSCo generates human-interpretable representations for *any* unseen input.

To provide evidence for or against the above hypotheses, we evaluate the objective and subjective metrics from the human-robot teaming tasks and questionnaires. The utilized metrics and their implications for the hypotheses are detailed in the following.

V. Results and Analysis

Our detailed experiments involved 21 participants, aged 18 to 36, including four females and seventeen males. The study received approval from the Institutional Review Board (IRB) under the ID STUDY00019583. We defined three independent variables for SiSCo's signal types: NLS, VSM, and VSIntPro.

H1: Task Performance

To evaluate task performance, we utilized **OM1** for the measurement of task accuracy and **OM2** to assess task efficiency, considering both as dependent variables. For analyzing differences in task success rates across different signal modalities, we treated the data from **OM1** as binary categorical measurements and applied Fisher's exact test to determine statistical significance. The results indicated a significant difference in success rates between VSM and NLI ($stat$ = 0.25, $p < 0.05$), and between VSIntPro and NLI ($stat$ = 0.17, $p < 0.05$). However, the difference between VSIntPro and VSM was not statistically significant ($stat$ = 0.7, $p > 0.05$). Despite this, data presented in Table II shows that VSIntPro outperformed NLI by 18% and VSM by 3% in the overall task success rate. To assess task efficiency, we analyzed comprehension time and task completion time based on **OM2** data to ascertain any significant differences. The distribution of the **OM2** data was found to be non-normal as confirmed by a normality test [35]. Consequently, we employed the non-parametric Kruskal-Wallis test, a counterpart to the one-way ANOVA suitable for non-normal distributions, applying the Bonferroni correction to the p-values to mitigate the risk of

TABLE II: Objective Metrics: Averaged across all participants

OM1: Teaming Task Success Rate (%)						
	Δ (Desc.)	Θ (Col.)	Φ (Pos.)	Ω (Inst.)	Π (Ori.)	All
NLS	97.6	100.0	73.8	66.7	73.8	82.4
VSM	97.6	97.6	92.9	92.9	90.5	94.3
VSIntPro	97.6	97.6	100.0	100.0	90.5	97.1

OM2: Teaming Task Temporal Analysis (s)		
	Comprehension Time	Completion Time
NLS	28.9 $\pm$ 14.6	42.9 $\pm$ 19.8
VSM	5.7 $\pm$ 2.8	15.5 $\pm$ 5.5
VSIntPro	3.7 $\pm$ 1.3	11.4 $\pm$ 3.0

OM3: Questionnaire Object Recognition Success Rate (%)						
	Tennis Ball	Apple	Bread	Robot	Orange	All
VSIntPro	100.0	68.4	100.0	94.7	87.5	90.1

TABLE III: Subjective Metrics: Averaged across all participants

SM1: NASA-TLX (0 to 20)			
	Mental Demand	Temporal Demand	Frustration
NLS	7.4 $\pm$ 4.9	4.3 $\pm$ 3.8	5.0 $\pm$4.6
VSM	4.7 $\pm$ 4.0	3.9 $\pm$ 3.1	3.2 $\pm$ 3.3
VSIntPro	2.9 $\pm$ 2.4	3.4 $\pm$ 3.1	2.7 $\pm$ 2.6

SM2: SUS (0 to 100)	**SM3:** Signaling Modes Ranking (3 to 1)		
SiSCo	VSIntPro	VSM	NLS
82.0	1.1 $\pm$ 0.5	2.0 $\pm$ 0.4	2.9 $\pm$ 0.3

SM4: Object Recognition Rating (-5 to 5)					
	Tennis Ball	Apple	Bread	Robot	Orange
VSIntPro	3.3 $\pm$ 2.0	3.2 $\pm$ 1.8	1.2 $\pm$ 2.9	4.3 $\pm$ 1.6	1.3 $\pm$ 2.8

SM5: Signal Re-Engineering	**SM6:** Creative Inputs	
Rating (-5 to 5)	Rating (-5 to 5)	
VSIntPro	2.3 $\pm$ 2.1	VSIntPro	2.2 $\pm$ 3.0

Type-I error. The results revealed significant differences in communication efficiency. In terms of comprehension time, a clear disparity was observed between VSIntPro and NLI (H = 56.1, $p < 0.05$), VSM and NLI (H = 52.6, $p < 0.05$), and VSIntPro and VSM (H = 15.6, $p < 0.05$). Similarly, the completion time exhibited a comparable trend, with significant differences between VSIntPro and NLI (H = 108.3, $p < 0.05$), VSM and NLI (H = 77.1, $p < 0.05$), and VSIntPro and VSM (H = 6.3, $p < 0.05$). These findings indicate that VSIntPro is the most effective communication mode among those evaluated. The aforementioned results show that SiSCo's visual signals significantly enhance task performance, thereby corroborating hypothesis **H1**. Furthermore, although VSIntPro did not significantly outperform VSM in task accuracy, participants ranked it as the superior mode of communication in the subjective questionnaire (**SM3** in Table III).

H2: Cognitive Load

In the human-robot teaming task, participants remained stationary on table, which led to the decision to omit the physical demand and effort components of the NASA-TLX. Furthermore, the performance subscale from NASA-TLX was excluded as participants could not evaluate their successful task completion at the conclusion of each experiment as only expert knows whether human participant succeeded or not. Thus, we concentrated on mental demand, temporal demand, and frustration as our dependent variables to assess cognitive load, which are documented as **SM1** shown in Table III. Statistical verification of data normality was conducted based on D'Agostino's tests [35], paving the way for the Multivariate Analysis of Variance (MANOVA). The results delineated significant cognitive load variations when comparing Visual Signals with Intention Projection (VSIntPro) and the Natural Language Interface (NLI) ($F(3,32)$ = 4.5, $p < 0.05$), as well as between Visual Signals with a Monitor (VSM) and NLI ($F(3,32)$ = 3.0, $p < 0.05$). However, the difference between VSIntPro and VSM was not statistically significant ($F(3,32)$ = 0.9, $p > 0.05$). On average, VSIntPro achieved 46% lower score on NASA-TLX subscales compared to NLI showing its prowess in easing the communication. In terms of effectiveness in conveying information, ranked by **SM3** (see Table III), VSIntPro attained the highest rank score (1.1 $\pm$ 0.5), followed by VSM (2.0 $\pm$ 0.4) and NLI (2.9 $\pm$ 0.3), indicating a preference among participants for VSIntPro as the most effective mode. With VSIntPro demonstrating the least cognitive load and achieving the highest effectiveness ranking, we confirm our hypothesis **H2**.

H3: Generalization

In our pursuit to enhance human-robot collaboration, assessing SiSCo's capacity for handling various human inputs was crucial. We assessed this adaptability by integrating subjective metrics (**SM4-6**) for users to evaluate the visual signals SiSCo produced in response to new inputs, alongside an objective metric (**OM3**) that quantified how accurately participants identified target objects.

Despite initial analyses, user ratings deviated from a normal distribution, failing to fulfill the criteria set by D'Agostino's tests [35]. Consequently, we employed the non-parametric Wilcoxon signed-rank test to determine if user ratings were significantly more favorable compared to a baseline rating of "0", which is hypothesized as a pivot for potential improvement in a range from 0 to 5.

The Wilcoxon test affirmed that user ratings were significantly above the baseline for all subjective metrics (**SM4**: $stat$ = 2976, $p < 0.05$; **SM5**: $stat$ = 181.5, $p < 0.05$; **SM6**: $stat$ = 147.5, $p < 0.05$), suggesting a decided preference for the SiSCo-generated visual signals. In conjunction with these findings, **OM3** revealed that participants successfully identified the correct objects in, on avarage, 90.1% of cases, illustrating SiSCo's reliability. Collectively, these results substantiate hypothesis **H3**, affirming both the user satisfaction with SiSCo and its effectiveness in a generalized application context.

Additionally, the System Usability Scale (SUS) was administered to gauge user perceptions of SiSCo's adaptability for human-robot collaboration. With a score of 82.0, SiSCo significantly surpassed the average SUS benchmark, reflecting a robust endorsement of the system's usability by the participants.

VI. Discussion and Limitations

We introduced SiSCo, a novel framework designed to enhance communication in human-robot collaboration by

integrating mixed-reality environments with LLMs. SiSCo generates visual signals and cues on the fly to convey robotic intent and to influence human intervention. In this work, we utilized GPT4-Turbo for signal generation, which causes a notable dependence on OpenAI's server infrastructure. As LLM research advances in the future, we will focus on fine-tuning a model for local execution to mitigate this limitation.

Our evaluation of SiSCo has revealed remarkable adeptness in generating signals tailored for facilitating human-robot collaboration tasks. Even when generating visual signals for previously unseen objects, users rated the object representation quality as above average in all subjective metrics. Objective metrics underscored these findings with a significant enhancement in measured task performance, 73% faster task completion time and 18% higher task success rate, when utilizing visual signals generated through intention projection (VSIntPro) compared to simple natural language signals. Additionally, users reported a substantial 46% reduction in cognitive load with the VSIntPro mode, as measured by NASA-TLX subscales. Taken together, the adeptness of SiSCo in signal generation and the high level of user satisfaction elicit a strong potential for future endeavors integrating LLMs and mixed reality in human-robot collaboration.

REFERENCES

[1] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, vol. 1, no. 4, pp. 279–285, 2020.

[2] R. S. Andersen, O. Madsen, T. B. Moeslund, and H. B. Amor, "Projecting robot intentions into human environments," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 294–301.

[3] S. H. Choi, K.-B. Park, D. H. Roh, J. Y. Lee, M. Mohammed, Y. Ghasemi, and H. Jeong, "An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102258, 2022.

[4] L. F. González-Böhme and E. Valenzuela-Astudillo, "Mixed reality for safe and reliable human-robot collaboration in timber frame construction," *Buildings*, vol. 13, no. 8, p. 1965, 2023.

[5] A. Rivera-Pinto, J. Kildal, and E. Lazkano, "Toward programming a collaborative robot by interacting with its digital twin in a mixed reality environment," *International Journal of Human–Computer Interaction*, pp. 1–13, 2023.

[6] P. Bottoni, M. Costabile, S. Levialdi, and P. Mussio, "Formalising visual languages," in *Proceedings of Symposium on Visual Languages*, 1995.

[7] R. K. Ganesan, Y. K. Rathore, H. M. Ross, and H. B. Amor, "Better teaming through visual cues: how projecting imagery in a workspace can improve human-robot collaboration," *IEEE R&A Magazine*, 2018.

[8] R. S. Lunding, M. S. Lunding, T. Feuchtner, M. G. Petersen, K. Grønbæk, and R. Suzuki, "Robovisar: Immersive authoring of condition-based ar robot visualisations," in *2024 ACM/IEEE HRI*, ser. HRI '24. ACM, 2024, p. 462–471.

[9] R. Raj, A. Singh, V. Kumar, and P. Verma, "Analyzing the potential benefits and use cases of chatgpt as a tool for improving the efficiency and effectiveness of business operations," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 3, p. 100140, 2023.

[10] S. Sok and K. Heng, "Chatgpt for education and research: A review of benefits and risks," *Available at SSRN 4378735*, 2023.

[11] M. Pascher, U. Gruenefeld, S. Schneegass, and J. Gerken, "How to communicate robot motion intent: A scoping review," in *Proceedings of the 2023 CHI CHFCS*, 2023, pp. 1–17.

[12] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[13] Y. Che, A. M. Okamura, and D. Sadigh, "Efficient and trustworthy social navigation via explicit and implicit robot–human communication," *IEEE Trans. on Robotics*, vol. 36, no. 3, pp. 692–707, 2020.

[14] C. Brosque, E. G. Herrero, Y. Chen, R. Joshi, O. Khatib, and M. Fischer, "Collaborativewelding and joint sealing robots with haptic feedback," in *ISARC*, vol. 38. IAARC Publications, 2021, pp. 1–8.

[15] M. A. Cabrera, J. Heredia, J. Tirado, V. Panov, F. Hagos, and D. Tsetserukou, "Cohaptics: Development of human-robot collaborative system with forearm-worn haptic display to increase safety in future factories," in *2021 IEEE CASE*. IEEE, 2021, pp. 74–80.

[16] P. Neto, M. Simão, N. Mendes, and M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," *IJAMT*, vol. 101, pp. 119–135, 2019.

[17] C. Carissoli, L. Negri, M. Bassi, F. A. Storm, and A. Delle Fave, "Mental workload and human-robot interaction in collaborative tasks: A scoping review," *IJHCI*, pp. 1–20, 2023.

[18] T. Williams, D. Szafir, T. Chakraborti, and H. Ben Amor, "Virtual, augmented, and mixed reality for human-robot interaction," in *Companion of the 2018 ACM/IEEE HRI*, 2018, pp. 403–404.

[19] E. Chang, H. T. Kim, and B. Yoo, "Virtual reality sickness: a review of causes and measurements," *International Journal of Human–Computer Interaction*, vol. 36, no. 17, pp. 1658–1682, 2020.

[20] S. Rokhsaritalemi, A. Sadeghi-Niaraki, and S.-M. Choi, "A review on mixed reality: Current trends, challenges and prospects," *Applied Sciences*, vol. 10, no. 2, p. 636, 2020.

[21] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson, "Visualizing big data with augmented and virtual reality: challenges and research agenda," *Journal of Big Data*, vol. 2, no. 1, pp. 1–27, 2015.

[22] P. S. Dunston and X. Wang, "Mixed reality-based visualization interfaces for architecture, engineering, and construction industry," *Journal of construction engineering and management*, vol. 131, no. 12, pp. 1301–1309, 2005.

[23] T. Wu, M. Terry, and C. J. Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–22.

[24] G. Kim, P. Baldi, and S. McAleer, "Language models can solve computer tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[25] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[26] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[28] S. Sonawani, Y. Zhou, and H. B. Amor, "Projecting robot intentions through visual cues: Static vs. dynamic signaling," in *2023 IEEE/RSJ IROS*. IEEE, 2023, pp. 7931–7938.

[29] ——, "Imitation learning based auto-correction of extrinsic parameters for a mixed-reality setup," in *1st XR-ROB Workshop, IROS 2022*, 2022.

[30] S. Sonawani and H. Amor, "When and where are you going? a mixed-reality framework for human robot collaboration," in *VAM - HRI*, 2022.

[31] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[32] https://platform.openai.com/docs/api-reference/, [Acc. Feb-2024].

[33] M. Feick, N. Kleer, A. Tang, and A. Krüger, "The virtual reality questionnaire toolkit," in *ACM Symposium on User Interface Software and Technology*, 2020.

[34] J. Brooke, "Sus: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, 11 1995.

[35] R. D'agostino and E. S. Pearson, "Tests for departure from normality. empirical results for the distributions of b 2 and− b," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.