Reimagining RViz: Multidimensional Augmented Reality Robot Signal Design

Thomas R Groechel*, Amy O'Connell*, Massimiliano Nigro*, and Maja J Matarić

***Abstract*— From RViz to augmented reality (AR), a wide variety of robot signal visualizations exist for conveying robot capabilities. Many of the visualizations designed for AR, however, have not isolated multiple salient Virtual Design Elements (VDEs) for a given signal and comparatively evaluated combinations of those VDEs. To address this, we identify multiple VDEs for AR signaling of the following core robot capabilities: *navigation*, *light detection and ranging (LiDAR)*, *camera*, *face detection*, *audio localization*, and *natural language processing*. We evaluated each signal's VDE combinations with an Amazon Mechanical Turk study (n=150) where participants watched 4 videos for each signal (consisting of 2 independent VDE choices) and rated the clarity and visual appeal of each signal. The results define a set of the most clear and visually appealing signal visualization designs and inform about interaction effects among VDEs. The resulting VDEs offer design insights and a baseline for continued research into AR robot capability signalling.**

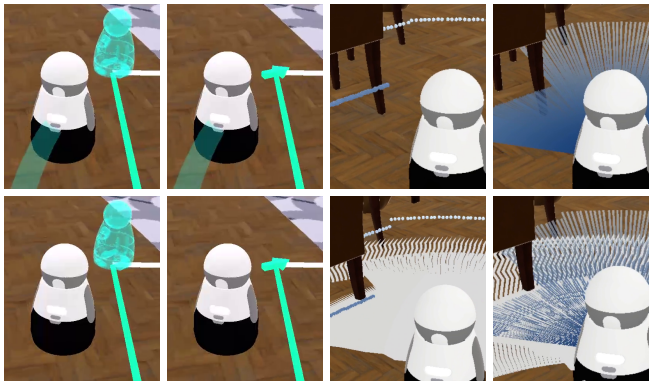


Fig. 1: Combinations of Virtual Design Elements (VDEs) for navigation visualization (left) and LiDAR visualization (right). Details of each signal design are found in Sec. III.

I. INTRODUCTION

Whether it is a young student interacting with a socially assistive robot tutor in school or a trained roboticist debugging a system being created, the ability for a user to accurately estimate a robot's capabilities is critical for effective human-robot interaction (HRI). Within HRI research, *perceived robot capability* is defined as a user's perception of the robot's true capabilities [1]. Under- and over-perception refer to a mismatch between the user's perceptions about the robot and the robot's true capabilities [1]. Over-perception occurs when a user expects the robot to do something it is not capable of, while under-perception occurs when the user misses an interaction capability the robot possesses.

Visualizing robot capabilities as explicitly as possible aids capability signalling. Widely used software such as RViz [2] displays robot sensors (e.g., cameras) and reasoning capabilities (e.g., mapping and navigation waypoints). Further situating such visualizations, the field of virtual, augmented, and mixed reality for human-robot interaction (VAM-HRI [3]) creates virtual objects and places them in a 3D VR space or directly projects virtual imagery onto the real world. VAM-HRI aims to increase the robot's *Expressivity of View* (EV) [4], the ability to express its capabilities and its internal *Complexity of Model* (CM) [5]. There are many examples of increasing EV in VAM-HRI [6], from recreating RViz for AR (e.g., iviz [7]) to creating VR interfaces [8].

Many of these VAM-HRI signal designs, however, either explore a distinct set of visualizations (e.g., [9]) or only pose a single signal design for comparison against a non-VAM interface (e.g., RViz). To address these shortcomings, we created a set of salient Virtual Design Elements (VDEs [6]) for a set of core robot capabilities: **navigation**, **light detection and ranging (LiDAR)**, **camera**, **face detection**, **audio localization**, and **natural language processing (NLU)**. We created two independent VDEs for each capability and validated the pairwise designs on Amazon Mechanical Turk (AMT) in a video-based study (n=150). AMT is a standard tool used in VAM-HRI research for initial signal design studies [5]. Four videos of each signal displaying all combinations of the independent VDEs were shown to participants and evaluated for clarity and visual appeal. The results determine the clearest and most visually appealing design choices while also highlighting possible interaction effects of intra-signal VDEs.

II. BACKGROUND

A. Perceiving robot capabilities

Understanding a given robot's capabilities is increasingly important as robots become more common in everyday life [**?**]. Currently, robots are often incomprehensible to users, who struggle to predict their capabilities and intentions due to poor information exchange where neither users nor robots can understand what the other is explicitly or implicitly communicating [6]. Several studies have aimed to improve human-robot communication by having the robot understand

*Equal Contribution

This work was supported by the NSF NRI 2.0 grant for "Communicate, Share, Adapt: A Mixed Reality Framework for Facilitating Robot Integration and Customization", NSF IIS-1925083.

Thomas Groechel, Amy O'Connell, Massimiliano Nigro, and Maja J Matarić are all with the Interaction Lab, Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA
{groechel,amy.dell,mn_902,mataric}@usc.edu

and communicate through human non-verbal cues such as gaze, gesture, and natural language [10]. Such communication methods, however, may not be sufficient to convey complex information and are not generalizable to all robot embodiments; communication should be form-agnostic in its ability to convey complex robot signals [10].

B. Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)

The VAM-HRI field focuses on leveraging robot-agnostic VAM technologies (e.g., VR/AR head-mounted displays) to facilitate communication. VAM technology can be used independently of the robot's embodiment and can communicate complex signals through 3D virtual imagery [6]. Such imagery has been shown to increase the ease of robot programming, remote teleoperation [7], [8], [11], human intent estimation [12], socially assistive tasks [13], and human-robot teaming tasks [9]. AR visualizations of robot signals improve human-robot communication by providing complex information in a simplified and accessible manner [6].

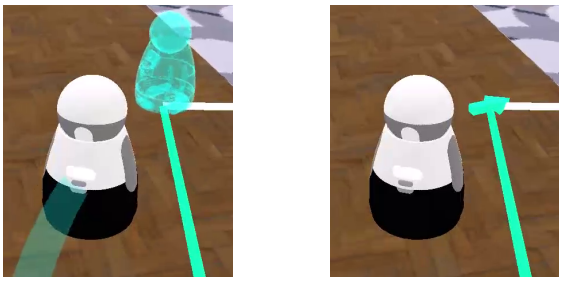
VAM-HRI works focusing on increasing a robot's Expressivity of View [5], however, typically create a task-specific set of distinct signals (e.g., [9]) or create a single design to be compared against a traditional interface (e.g., [8]) or no interface (e.g., [13]). To address this limitation, our work explores the design space for 6 common robot capabilities by pairing and comparing multiple Virtual Design Elements (VDEs [6]) for each signal. The goal of this work is to better understand the design space for VAM-HRI robot capability signalling.

III. VIRTUAL DESIGN ELEMENTS OF A SIGNAL

To create each signal's Virtual Design Element (VDE [6]), we took inspiration from prior visualization research and existing visualization software (e.g., RViz [2]). We used Unity 3D game engine v2021.2.7f1 to create the visualizations for this work and have made them open-source and available at `https://github.com/interaction-lab/NRI-SVTE`. A video of all signal VDE combinations we created can be found at `https://youtu.be/Xw2_kHyN-xA`.

A. Navigation

We created visualizations for the robot's ability to plan and execute paths in the environment.



(a) Robot Outline + Trail (b) Arrow + No Trail

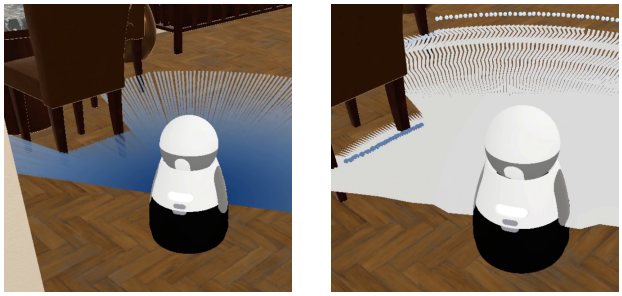
Fig. 2: Navigation visualizations.

Robot Outlines / Arrows: Two different sets of navigation waypoints were designed: ghost robot outlines and arrows. The ghost robot outlines (Fig. 7a), inspired by [8], place semi-transparent robots at each waypoint, oriented in the goal direction of the waypoint. The outlines were scaled down so as to reduce visual clutter and allow the robot to fully cover the outline. The 3D arrows were inspired by RViz [2]. We expected the differences between the two to be in the salience of direction (arrows $>$ outline), environment occlusion (arrow $>$ outline), and visual appeal (outline $>$ arrow).

Trail / No Trail: For a sense of momentum and direction, a trail was added to the visualization (Fig. 7a), but it occludes some of the environment.

B. LiDAR

We created visualizations of the robot's ability to use LiDAR sensors to detect nearby objects.



(a) Lines + No Pulse (b) Points + Pulse

Fig. 3: LiDAR visualizations.

Lines / Points: The robot measures the distance to objects in the environment with a 180 LiDAR sensors spaced 1° apart. 3D lines and points were used to indicate the distance returned by each of the sensors. The lines (see Fig. 3a) occupied the same position as the laser beam emitted by each sensor and were displayed as opaque and colored, with a gradient from dark blue at the source to white at the maximum distance measurable by the sensor. The points, inspired by RViz [2], were opaque 3D spheres. The color of each sphere was modulated to indicate the distance from the robot according to the same gradient scale used on the lines. The expected differences were in the salience of distance from the robot (lines $>$ points), indication of lasers (lines $>$ points), environment occlusion (points $>$ lines) and visual appeal (lines $>$ points).

Pulse / No Pulse: In order to give the impression that the sensors emitted lasers to measure distance, fixed length white lines were visualized emitting at a fixed interval from each sensor. The lines occluded part of the environment and do not continue to align with the sensors as the robot moves, visible in Fig. 3b.

C. Camera

We created visualizations of the robot's ability to to collect visual information through a camera located in its left eye.

(a) HUD + No Pyramid (b) In Scene + Pyramid

Fig. 4: Camera visualizations.

In Scene / HUD: Input from the robot's camera was visualized in one of two locations depending on the condition. In the in scene condition, the input was visible in a square window that hovered above the robot (see Fig. 4b). As the video rotated around the robot to show the scene from multiple viewpoints, the window rotated to always face the participant. In the HUD (heads-up display) condition, the input was visible in a fixed square on the screen. In both conditions, a grey frame was placed around the window to increase salience. These visualizations were inspired by the difference in "User-anchored" and "Robot-anchored" mixed reality design elements [5], [6].

Pyramid / No Pyramid: A 3D translucent pyramid (see Fig. 4b) visualized the portion of the scene that was visible to the robot's camera, inspired by camera depictions in game engines. The pyramid was fixed to the robot's eye containing the camera and rotated with the robot's head to correspond with the camera input at all points in the video. The expected differences were in the salience of the portion of the environment visible to the robot (pyramid > no pyramid), environment occlusion (no pyramid > pyramid), and how appealing it is to look at (pyramid > no pyramid).

D. Face Detection

We developed visualizations of the robot's ability to detect faces from its camera input.

(a) Box + In Scene (b) Mesh + Not In Scene

Fig. 5: Face detection visualizations.

Box / Face Mesh: Two different sets of face markers were designed: green boxes and triangle meshes. The boxes (see Fig. 5a), inspired by OpenCV [14] face detection software, were green square boxes fixed around the faces of each person in the robot's camera's field of view. Translucent triangle meshes, inspired by face detection software such as Google's MediaPipe [15], were placed on each character model in the same way as the boxes. Both face markers were sized to fit each character model and rotated with the robot's head to remain oriented toward the robot's camera.

In Scene / Not In Scene: The boxes and meshes were always visible in the camera input on the HUD. However, in the in scene condition, the face markers were also visualized as objects fixed to the faces of the people in the scene. In the **not in scene** condition, the markers were only visible in the HUD window but not in the environment.

The expected differences were in the salience of faces (in scene > not in scene) and environment occlusion (in scene > not in scene), inspired by the difference in "User-anchored" and "Environment-anchored" mixed reality design elements [5], [6].

E. Audio Localization

We created visualizations of the robot's ability to to estimate the positions of sources sound in the environment.

(a) Cones + Small (b) Spheres + Large

Fig. 6: Audio localization visualizations.

Spheres / Cones: 3D objects of two distinct shapes (spheres, cones) were overlayed around the robot and increased in size and color gradient in relation to the loudness input received from the directional microphones. If no sound was perceived, the objects remained hidden. The cones and the spheres were inspired by [16] and [17], [18], respectively.

Small / Large: Small/Large visualizations differed by how much the 3D objects were increasing in size with respect to the microphone input. This characteristic explored the trade-off between robot visibility and visual indication of loudness.

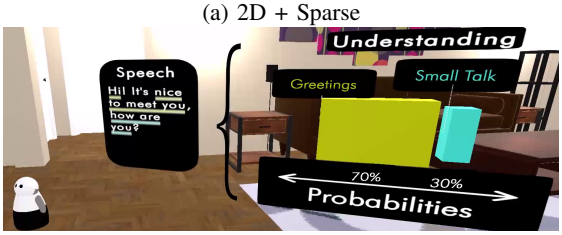
F. Natural Language Understanding (NLU)

We created visualizations of the robot's ability to to analyze spoken language and make predictions about the user's meaning. The main elements of the visualizations were:

- A speech bubble containing the speech understood by the robot in real time, with keyword highlighting [19] as soon as the intents were extracted by NLU;
- A horizontal bar diagram, indicating the confidence for each intent understood in the sentence;
- Graphical elements (e.g., labels, curly brace) providing additional contextual information about the elements in the visualization.

Cluttered / Sparse: We explored the effect of the decluttering principle [20] on our visualization. The cluttered version presented all the graphical elements, providing more

(a) 2D + Sparse



(b) 3D + Cluttered

Fig. 7: NLU visualizations.

information to the user but cluttering the visualization. The sparse version removed the graphical elements and provided less information about the relationship between the elements in the visualization.

3D / 2D: In the 3D/2D versions, the bar chart was displayed in its 3D or its 2D version.

IV. METHODS

To increase reproducibility, the study methods are also included in the open-source repository wiki `https://github.com/interaction-lab/NRI-SVTE/wiki`.

A. Participants

Participants were recruited through AMT. To determine the study size, we followed [21]: $50 + 8 * m$ where m is the number of independent variables (2 for each of the 6 signals, thus $m = 12$), resulting in $50+8*12 = 146$ participants. We added 4 more participants in the case of incomplete data.

Inclusion criteria for the study were:

- At least 18 years of age;
- In the United States or US Minor Outlying Islands;
- Number of AMT HITs Approved > 1000;
- AMT HIT Approval Rate $\geq$ 99%.

The 150 participants who completed the survey identified in open-ended questions as: Gender Identity – Cis Woman : 1, Female: 52, Male: 95, and left blank: 2; Race – African American : 3, American : 1, Asian : 8, Black : 9, Caucasian : 13, European : 1, Hispanic : 3, Latina : 1, Middle Eastern : 1, Native American : 1, White : 101, and left blank : 8; Age – Mdn : 35, $\bar{X}$: 37.22, σ : 9.74, and Range(20,71).

B. Procedure and Measures

This study was approved by the University's Institutional Review Board (IRB #UP-20-00030). Each participant first consented to the study, confirmed that their audio worked, and filled out a set of demographic questions. We used a within-subjects study design in which participants were shown four ten-second videos displaying each combination of conditions for the six signal visualizations described in Sec. III. Both the signal set order and the order of videos within each signal set were randomized, with the exception that the camera visualization was always shown immediately before the face detection visualization, as the latter depended on the former.

We collected quantitative and qualitative data to measure the participants' opinions of the visualizations and their understanding of the robot based on the videos (found at `https://youtu.be/Xw2_kHyN-xA`), shown from the perspective of someone wearing an AR headset. The camera moved to different points of view as the robot moved to demonstrate a capability. Camera views were replicated exactly within each signal across conditions.

1) Quantitative Data: Participants were not told what capability of the robot was visualized in the videos when they first watched them. For each video, they rated the clarity and visual appeal via two 7-point Likert items ("The video shows the capability in a clear way", "The video is visually appealing") from "Strongly Disagree" to "Strongly Agree." The items were adapted from [22] and [23], respectively.

2) Qualitative Data: After watching and rating the four videos, participants were asked the following three open-ended questions: "What did you like about the visualizations you scored higher?" "What did you dislike about the visualizations you scored lower?" and "What capability of the robot do you think is visualized in the videos above?". Once they had answered these three questions, the signal name and a short paragraph describing the signal (similar to those found in Sec. III) were revealed to the participant. They were then asked one additional open-ended question, "How could the visualizations above be improved to better illustrate the robot's ability to perform [signal name]?"

Participants were required to watch all four videos and answer all questions about them before moving on to the next signal. The process was repeated until all six signal sets were completed. Upon completion, each participant received an Amazon gift card worth US$6.25.

C. Analysis

We performed both quantitative and qualitative analyses of the results. Survey data were treated as ordinal so we used non-parametric tests [24]. To compare signal clarity and visual appeal, Wilcoxon signed-rank tests were used with Holm's corrected p values [25] and α levels <.05*, <0.1**, and <.001***. The brute force common language effect size (CLES) proposed by Varga et al. [26] was calculated; mean and standard deviation calculations are not appropriate for ordinal data [24]. CLES is the proportion of paired samples (s_{G0}, s_{G1}) where s_{G0} is higher than s_{G1}. To confirm the signals portrayed the intended capability, we also annotated the open response question "What does this visualization portray?" with the criteria set in Sec. III. For qualitative analysis, we read through all participant open-ended answers marking down themes and associated quotes to each theme. All 150 participants were included in the analysis given the strict criteria described in Sec. IV-A.

TABLE I: Navigation results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
RobotOutline + Trail / Arrow + No Trail	0.614	<.001***
RobotOutline + No Trail / Arrow + No Trail	0.589	<.001***
RobotOutline + Trail / Arrow + Trail	0.568	.015*
Arrow + Trail / Arrow + No Trail	0.549	.312
RobotOutline + No Trail / Arrow + Trail	0.543	.449
RobotOutline + Trail / RobotOutline + No Trail	0.526	.58
Visual Appeal		
RobotOutline + No Trail / Arrow + No Trail	0.589	.001**
RobotOutline + Trail / Arrow + No Trail	0.598	.003**
RobotOutline + Trail / Arrow + Trail	0.565	.178
RobotOutline + No Trail / Arrow + Trail	0.555	.25
Arrow + Trail / Arrow + No Trail	0.538	.427
RobotOutline + Trail / RobotOutline + No Trail	0.51	.999

TABLE II: LiDAR results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Line + No Pulse / Point + No Pulse	0.658	<.001***
Point + Pulse / Point + No Pulse	0.598	<.001***
Line + No Pulse / Line + Pulse	0.586	<.001***
Line + No Pulse / Point + Pulse	0.56	.036*
Line + Pulse / Point + No Pulse	0.563	.098
Point + Pulse / Line + Pulse	0.528	.422
Visual Appeal		
Line + No Pulse / Line + Pulse	0.662	<.001***
Line + No Pulse / Point + Pulse	0.622	<.001***
Point + No Pulse / Line + Pulse	0.595	<.001***
Line + No Pulse / Point + No Pulse	0.584	.003**
Point + No Pulse / Point + Pulse	0.55	.064
Point + Pulse / Line + Pulse	0.541	.134

V. RESULTS

A. Navigation

1) Quantitative: Survey results for Wilcoxon signed-rank test are given in Table I. A total of 102 participants (68%) correctly identified the robot's navigation capability.

2) Qualitative Analysis: Likes/Dislikes– Participants reported liking the arrows for their simplicity, clarity of direction, and lack of clutter, reporting the opposite for the robot outlines. Alternatively, other participants enjoyed the robot outlines citing them as more clear and visually "exciting" or "cool", reporting the arrows as "boring". A similar trend emerged for robot trails where participants liked the clarity of the trail (e.g., **P12**: "I like the trial in which it left behind to show where it was coming from and how it was moving with the way points") while those opposed disliked the clutter (e.g., **P67**: "I didn't like the blue line that followed him showing him actually doing it. It was overkill").

Suggested improvements– A common theme of dynamic objects and colors were suggested (e.g., **P12** : "as the robot goes through the path, the way point that the robot travel should fade or change colors"). Another suggested theme were requests for indication of robot planning (e.g., **P96**: "Before the robot moves, there should be a small thought bubble with a travel plan") as well as a map (e.g., **P100**: "Show a map").

Other unique factors– Participants described the visualization videos as "smoother" than others (e.g., **P1**: "[It was] less busy and the camera less nauseating"). Participants also reported confusion as to why the robot did not follow the straight line path (e.g., **P138**: "Why did it 'walk' in a rounded fashion when the lines were straight?").

B. LiDAR

1) Quantitative Analysis: Survey results for Wilcoxon signed-rank test are shown in Table II. A total of 62 participants ($\approx 41.3\%$) correctly identified the robot's LiDAR capability.

2) Qualitative Analysis: Likes/Dislikes– Rather than commenting on the individual points or lines for each sensor, participants emphasized the overall shape they formed when combined. For example, participants perceived the points as forming a "line" to denote the boundary around the area that the robot could sense, whereas they described the lines as forming a "cone" or "fan" around the front of the robot. Some preferred the points because they left the area around the robot visible (e.g., **P131**: "The outline of the sensor area was clean and simple") while others found the points too simplistic to convey the relevant information (e.g., **P73**: "Just having a line is confusing and doesn't really tell you much."). Participants described the pulsing videos as "confusing," "laggy," and "glitchy" (e.g., **P144**: "It looks like a glitch, as if the robot is creating ice or something when it scoots along.") although some understood the pulsing to be a more accurate representation of how the data are collected.

Suggested improvements– Participants suggested somehow indicating the objects that had been detected (e.g., **P119** : "When approaching an object, the object should be shown in red for clarification."). Another theme among the suggestions was some indication of the laser light returning to the robot after reflecting off of an object (e.g., **P51**: "Maybe if the line bounced back like a radar").

Other unique factors– Participants suggested that these visualizations could be enhanced by including more textual information about what the robot is doing (e.g., **P60**: "Maybe include some sample distances that were sensed for the viewer to see more clearly what is going (for those numerically inclined individuals)").

C. Camera

1) Quantitative Amalysis: Survey results for Wilcoxon signed-rank test are shown in Table III. A total of 85 participants ($\approx 56.7\%$) correctly identified the camera capability.

2) Qualitative Analysis: Likes/Dislikes– Participants reported finding the pyramid helpful and correctly interpreted it as the portion of the scene visible to the robot (e.g., **P90**: "I liked that the visualization is able to show the field of vision for the robot and gave us a view of what the robot is seeing without blocking our vision."). Participants disliked that the pyramid distorted or obstructed the scene (e.g., **P72**: "I didn't like the way the colors muted when showing the

TABLE III: Camera results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Pyramid + HUD / No Pyramid + InScene	0.735	<.001***
Pyramid + HUD / No Pyramid + HUD	0.704	<.001***
Pyramid + InScene / No Pyramid + InScene	0.71	<.001***
Pyramid + InScene / No Pyramid + HUD	0.679	<.001***
No Pyramid + HUD / No Pyramid + InScene	0.54	.22
Pyramid + HUD / Pyramid + InScene	0.517	.999
Visual Appeal		
Pyramid + InScene / No Pyramid + InScene	0.621	<.001***
Pyramid + HUD / No Pyramid + InScene	0.608	<.001***
Pyramid + InScene / No Pyramid + HUD	0.597	.002**
Pyramid + HUD / No Pyramid + HUD	0.583	.005**
No Pyramid + HUD / No Pyramid + InScene	0.529	.999
Pyramid + InScene / Pyramid + HUD	0.518	.999

TABLE IV: Face detection results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Box + InScene / Mesh + InScene	0.756	<.001***
Box + InScene / Mesh + Not InScene	0.747	<.001***
Box + Not InScene / Mesh + InScene	0.738	<.001***
Box + Not InScene / Mesh + Not InScene	0.729	<.001***
Box + InScene / Box + Not InScene	0.524	.935
Mesh + Not InScene / Mesh + InScene	0.51	.999
Visual Appeal		
Box + Not InScene / Mesh + InScene	0.614	<.001***
Box + Not InScene / Mesh + Not InScene	0.585	.002**
Box + InScene / Mesh + InScene	0.584	.039*
Box + InScene / Mesh + Not InScene	0.558	.319
Box + Not InScene / Box + InScene	0.519	.999
Mesh + Not InScene / Mesh + InScene	0.529	.999

boundaries." **P56**: "The extra visual goodies kind of got in the way").

Participants preferences for the HUD or in-scene placement of the camera input were often related to the relative size of the displays. They found that the larger HUD display was easier to see (e.g., **P51**: "I liked that the inset picture was big enough to see well") but also obstructed the scene. They found the smaller in-scene window harder to see but less intrusive (e.g., **P84**: "I found the smaller window showing the family less intrusive"), although some still found the in-scene placement obstructive and misleading (e.g., **P12**: "The robots perspective being above its head felt like it blocked out the people, as well made it look like the robot was thinking of something").

Suggested improvements– Participants suggested several ways to make the pyramid less intrusive in the scene without decreasing the salience of the visualization. They involved finding other ways to maintain the outline of the camera frame without distorting the scene with the translucent material (e.g., **P59**: "Instead of having the lens view be transparent have it just show an outline or bounding box", **P72**: "May a border line where the boundaries would be, but keep the colors and visuals not as distorted").

Other unique factors– As observed in the other signals, participants assumed the robot could perform functions beyond what the real robot was capable of, and suggested the visualization could better indicate what the robot is not doing in addition to what it is. In this example, participants were unsure if the robot was merely "seeing" the people in real time, or if the robot was taking pictures or recording videos of the people pictured (e.g., **P120**: "I think that placing a small red light bulb that turns on when the robot is recording or capturing images").

D. Face Detection

1) Quantitative: Survey results for Wilcoxon signed-rank test are in Table IV. A total of 109 participants ($\approx 72.7\%$) correctly identified the robot's face detection capability.

2) Qualitative: Likes/Dislikes– Participants overwhelmingly viewed visualizations with green boxes around faces more favorably than those featuring triangle face meshes. Commonly cited reasons included visual salience (e.g., **P51**: "The green boxes on the faces are easy to see"), and more obvious meaning (e.g., **P63**: "The screen mapping of the face structure wasn't very clear and was a bit creepy"). The face meshes were often associated with mistrust of the robot, although both markers were described by some participants as unsettling (e.g.,**P15**: "I didn't like the green boxes around the heads. It felt like they were a target", **P134**: "The crosshairs on the faces wasn't necessary and looked weird and technologically scary").

Participants who preferred having the markers in the scene in addition to on the HUD reported it as a more obvious indication of what the robot was doing (e.g., **P14**: "I liked the green boxes on both the floating frame and the actual scene, as it was a good way to reference exactly what the robot was seeing and interpreting"), while others found the in-scene markers unnecessary (e.g.,**P56**: "It was clear what the robot was doing, without the green frames getting in the way in the main physical scene").

Suggested improvements– Participants suggested changing the visualizations to indicate that the robot could differentiate between people, by numbering the boxes, using different colored boxes for the different faces, or labeling the boxes with the names if the robot recognizes specific individuals. These suggestions are evidence of a larger trend of participants seeking to clarify the extent of the robot's capabilities through the visualizations (e.g., can the robot recognize human faces, or simply detect them?).

Other unique factors– Participants had mixed suggestions on what color the box should be. For some, the green boxes were familiar and recognizable, while others found the green "creepy" and suggested using a more neutral color.

E. Audio Localization

1) Quantitative Analysis: Survey results for Wilcoxon signed-rank test are shown in Table V. A total of 58 participants ($\approx 58.7\%$) correctly identified the robot's audio localization capability.

TABLE V: Audio localization results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Sphere + Small / Sphere + Large	0.523	.191
Cone + Large / Sphere + Large	0.532	.468
Cone + Large / Cone + Small	0.513	.999
Cone + Large / Sphere + Small	0.51	.999
Cone + Small / Sphere + Large	0.519	.999
Sphere + Small / Cone + Small	0.502	.999
Visual Appeal		
Cone + Large / Sphere + Large	0.554	.009**
Cone + Large / Sphere + Small	0.544	.219
Cone + Small / Sphere + Large	0.536	.338
Cone + Large / Cone + Small	0.52	.999
Cone + Small / Sphere + Small	0.524	.999
Sphere + Small / Sphere + Large	0.512	.999

TABLE VI: NLU results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Cluttered + 3D / Sparse + 3D	0.536	.092
Cluttered + 3D / Sparse + Flat	0.546	.245
Cluttered + 3D / Cluttered + Flat	0.52	.749
Cluttered + Flat / Sparse + Flat	0.527	.999
Cluttered + Flat / Sparse + 3D	0.517	.999
Sparse + 3D / Sparse + Flat	0.51	.999
Visual Appeal		
Cluttered + Flat / Sparse + Flat	0.526	.325
Cluttered + Flat / Cluttered + 3D	0.518	.783
Cluttered + Flat / Sparse + 3D	0.532	.999
Cluttered + 3D / Sparse + Flat	0.509	.999
Sparse + Flat / Sparse + 3D	0.507	.999
Cluttered + 3D / Sparse + 3D	0.517	.999

2) Qualitative: Likes/Dislikes– Participants who preferred the cones described them as cleaner (e.g., **P81**: "I liked the smaller circles visuals more because it looked cleaner") and less obstructive than the spheres (e.g.,**P101**: "The bubbles were too large and surrounded the robot"). Participants who preferred the spheres found them more salient (e.g., **P64**: "The bigger bubble made it clear where the robot was perceiving the noise from,"). Participants varied widely in their interpretations of the shapes, referring to the cones as "dots" and "arrows" and the to the spheres as "bubbles" or a "ring" formed around the robot as they overlapped. In some cases, these impressions contributed to the participants' interpretations of the visualizations (e.g., **P43** interpreted the cones as arrows and expected them to point at the source of the audio: "I didn't find the arrows very helpful as it did not accurately point to the source").

Suggested improvements– The most common suggestions included adding a compass or one or more arrows that point to the source of the sound to more clearly indicate that the robot is interested in the direction/location of the sound source, and displaying some kind of icon near the robot to more clearly signal that the robot can detect sound, such as an ear or microphone. Participants also suggested changing the sound source, currently a boom box animated to move to different locations around the robot in the environment, to something a robot is more likely to encounter in a home (e.g.,**P3**: "I think it can be improved by showing sound moving from something more realistic, maybe like a toy train going around a toy track with the robot in the middle").

Other unique factors– Participants were generally unclear about where the noise was coming from (i.e., some assumed the music was coming from the robot rather than the boom box, or that the robot was controlling the boom box).

F. Natural Language Understanding (NLU)

1) Quantitative: Survey results for Wilcoxon signed-rank test are shown in Table VI. A total of 112 participants ($\approx 74.7\%$) correctly identified the robot's NLU capability.

2) Qualitative Analysis: Likes/Dislikes– In their open-ended responses, participants expressed preferring cluttered and sparse displays in approximately equal measure, which is consistent with our analysis of the quantitative ratings. Participants who preferred cluttered displays valued the additional information and context provided by the additional axes and labels (e.g., **P56**: "The word "probabilities" helped to make clear what the robot was showing."). Conversely, participants who preferred sparse displays described non cluttered displays as "cleaner" and less confusing to look at (e.g.,**P16**: "They were less cluttered with labels than the ones that had the understanding and probability labels", **P102**: "The ones I scored lower had too much going on and made it a bit confusing"). Participants cited mainly aesthetic reasons for their ratings of the 3D and 2D graphs (e.g.**P14**: "I liked the bar graphs that were 3d, since they fit the overall aesthetic of the video better").

Suggested improvements– Participants indicated that the videos should be longer and show more interaction. Specifically, they formed the assumption that the robot talked back to the human, and were interested in observing how the robot would form a response based on the information it collected (e.g., **P12**: "Showing a response from the robot would help understand how it decides what response it takes"). They also suggested displaying more information about the robot's "thought processes" (e.g., **P14**: "It would be helpful to know which specific words are given more weight than others in determining which meaning. I also have to imagine some words would be classified into both categories, which would be helpful to be able to see visually").

Other unique factors– Participants hypothesized about other kinds of speech the robot understands and how the robot might act on its interpretations. Their suggestions involved testing the robot's knowledge or showing how it functions in varied situations (e.g., **P10**: "Could there be a graph for the robots understanding of body language aside from verbal communication?", **P122**: "Ask more question to test its knowledge base").

VI. DISCUSSION

This work presented Virtual Design Elements (VDEs) for 6 robot signals, using an AMT study to validate their designs

for clarity and visual appeal. Some VDEs consistently scored higher than others, such as visualizations that employed the pyramid for the camera and the box design over the mesh for face detection. In some other cases, there was not a consistently better VDE: different groups of users preferred different VDEs. In LiDAR visualizations, participants who preferred dots also preferred the trail due to the fan shape, while those who preferred lines did not want the trails due to the clutter. In navigation visualizations, a group of participants preferred fewer visual stimuli (e.g., those who preferred arrows + no trail).

Common signal-agnostic themes emerged in qualitative responses for the salience of the visual given the background as well as including more textual or verbal information. For salience, many participants cited VDEs disappearing into the background or being too small (e.g., face meshes). Enlarging the objects in the scene to make them more salient, however, does not appeal to everyone. A possible solution is making visualizations adapt to the background and change colors dynamically to improve saliency.

Further, participants suggested that we include both a verbal and text-based description of the scene to make things clearer (i.e., not relying on the visualizations alone). But introducing more objects in the scene would also introduce clutter. To make visualizations clear for all types of users they have to adapt to different preferences.

We provided the baseline for the design of VDEs for different signals and identified groups of users based on visual stimuli. We hope in the future to work on finding characteristics to identify user groups and to adapt the visualizations to different user groups. These visualizations serve as a basis for future signal design work.

VII. Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are undercited relative to the number of papers in the field. We recognize this bias and have worked diligently to ensure that we are referencing appropriate papers with fair gender and racial author inclusion. Please see the SVTE wiki for more information (linked in Sec. IV).

References

[1] E. Cha, A. D. Dragan, and S. S. Srinivasa, "Perceived robot capability," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 541–548, IEEE, 2015.

[2] "rviz." `https://github.com/ros-visualization/rviz`.

[3] C. T. Chang, E. Rosen, T. R. Groechel, M. Walker, and J. Z. Forde, "Virtual, augmented, and mixed reality for hri (vam-hri)," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1237–1240, 2022.

[4] T. Williams, D. Szafir, and T. Chakraborti, "The reality-virtuality interaction cube," in *Proceedings of the 2nd International Workshop on Virtual, Augmented, and Mixed Reality for HRI*, 2019.

[5] T. Groechel, M. Walker, C. T. Chang, E. Rosen, and J. Forde, "A tool for organizing key characteristics of virtual, augmented, and mixed reality for human-robot interaction systems: Synthesizing vam-hri trends and takeaways," *IEEE Robotics & Automation Magazine*, 2022.

[6] M. Walker, T. Phung, T. Chakraborti, T. Williams, and D. Szafir, "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy," *arXiv preprint arXiv:2202.11249*, 2022.

[7] A. Zea and U. D. Hanebeck, "iviz: A ros visualization app for mobile devices," *Software Impacts*, vol. 8, p. 100057, 2021.

[8] G. LeMasurier, J. Allspaw, M. Wonsick, J. Tukpah, T. Padir, H. Yanco, and E. Phillips, "Designing a user study for comparing 2d and vr human-in-the-loop robot planning interfaces," 2022.

[9] M. Walker, H. Hedayati, J. Lee, and D. Szafir, "Communicating robot motion intent with augmented reality," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 316–324, ACM, 2018.

[10] E. Cha, Y. Kim, T. Fong, M. J. Mataric, *et al.*, "A survey of nonverbal signaling methods for non-humanoid robots," *Foundations and Trends® in Robotics*, vol. 6, no. 4, pp. 211–323, 2018.

[11] E. Rosen, D. Whitney, E. Phillips, D. Ullman, and S. Tellex, "Testing robot teleoperation using a virtual reality interface with ros reality," in *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, pp. 1–4, 2018.

[12] E. Rosen, D. Whitney, E. Phillips, G. Chien, J. Tompkin, G. Konidaris, and S. Tellex, "Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays," vol. 38, no. 12, pp. 1513–1526. Publisher: SAGE Publications Ltd STM.

[13] T. Groechel, Z. Shi, R. Pakkar, and M. J. Matarić, "Using socially expressive mixed reality arms for enhancing low-expressivity robots," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–8, IEEE, 2019.

[14] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[16] Y. Kataoka, W. Teraoka, Y. Oikawa, and Y. Ikeda, "Effect of handy microphone movement in mixed reality visualization system of sound intensity," p. 8.

[17] A. Kose, A. Tepljakov, S. Astapov, D. Draheim, E. Petlenkov, and K. Vassiljeva, "Towards a synesthesia laboratory: Real-time localization and visualization of a sound source for virtual reality applications," vol. 14, no. 1, pp. 112–120. Number: 1.

[18] O. Lopez-Rincon and O. Starostenko, "Music visualization based on spherical projection with adjustable metrics," *IEEE Access*, vol. 7, pp. 140344–140354, 2019.

[19] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz, "Resilient chatbots: Repair strategy preferences for conversational breakdowns," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Association for Computing Machinery.

[20] K. Ajani, E. Lee, C. Xiong, C. N. Knaflic, W. Kemper, and S. Franconeri, "Declutter and focus: Empirically evaluating design guidelines for effective data communication," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[21] C. W. VanVoorhis, B. L. Morgan, *et al.*, "Understanding power and rules of thumb for determining sample sizes," *Tutorials in quantitative methods for psychology*, vol. 3, no. 2, pp. 43–50, 2007.

[22] A. Svalina, J. Pibernik, J. Dolić, and L. Mandić, "Data visualizations for the internet of things operational dashboard," in *2021 International Symposium ELMAR*, pp. 91–96, IEEE, 2021.

[23] F. Amini, N. H. Riche, B. Lee, J. Leboe-McGowan, and P. Irani, "Hooked on data videos: assessing the effect of animation and pictographs on viewer engagement," in *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, 2018.

[24] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 43–52, 2020.

[25] H. Abdi, "Holm's sequential bonferroni procedure," *Encyclopedia of research design*, vol. 1, no. 8, pp. 1–8, 2010.

[26] A. Vargha and H. D. Delaney, "A critique and improvement of the cl common language effect size statistics of mcgraw and wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.