VID2REAL HRI: Align video-based HRI study designs with real-world settings

Elliott Hauser[1], Yao-Cheng Chan[1], Sadanand Modak[2], Joydeep Biswas[2], Justin Hart[2]

Abstract—HRI research using autonomous robots in real-world settings can produce results with the highest ecological validity of any study modality, but many difficulties limit such studies' feasibility and effectiveness. We propose VID2REAL HRI, a research framework to maximize real-world insights offered by video-based studies. The VID2REAL HRI framework was used to design an online study using first-person videos of robots as real-world encounter surrogates. The online study ($n = 385$) distinguished the within-subjects effects of four robot behavioral conditions on perceived social intelligence and human willingness to help the robot enter an exterior door. A real-world, between-subjects replication ($n = 26$) using two conditions confirmed the validity of the online study's findings and the sufficiency of the participant recruitment target ($n = 22$) based on a power analysis of online study results. The VID2REAL HRI framework offers HRI researchers a principled way to take advantage of the efficiency of video-based study modalities while generating directly transferable knowledge of real-world HRI. Code and data from the study are provided at vid2real.github.io/vid2realHRI.

I. INTRODUCTION

Human-Robot Interaction (HRI) research is intimately shaped by the nature of robots as technical artifacts. As robots are not part of daily life for most people around the world, studies of robots in real-world settings are thus limited to existing uses of robots or researcher-constructed scenarios, which are oriented towards validating new designs or robot behaviors [1], and often involve real-world designs where humans are state-of-the-art research robots. While these may gain direct applicability to expanding the frontiers of robotics, they often lose ecological validity [2]. The difficulties this dynamic presents call for a new research framework that is applicable to novel forms of autonomous robot HRI phenomena, such as fully autonomous robotics and incidental human-robot encounters (HRE) research in pedestrian settings.

HRI of autonomous robots in real-world settings presents the methodological challenge of deriving precise knowledge of a highly complex subject in a high-dimensional space of potential variables. Incidental encounters with robots, the scenario studied here, are fleeting yet increasingly common events with potentially profound social impact. Without directly addressing these challenges, the field of HRI risks delaying or forgoing

the technical [3] and theoretical [4] opportunities of better understanding and reacting to them.

The present study proposes a new research framework, VID2REAL HRI, to guide researchers studying HRI with autonomous robots in real-world settings seeking to use video-based studies. VID2REAL HRI, described in Section III, allows researchers to better align video-based study designs with the real-world conditions they seek to understand.

We illustrate the application of the VID2REAL HRI framework to an HRE research study on the effects of socially compliant autonomous robot behaviors on pedestrians. This knowledge on pedestrian perceptions is crucial in evaluating robot behaviors when deployed for executing certain tasks while complying with social norms and preferences [5], [6]. The results of the application study demonstrated that VID2REAL HRI improved the video study's alignment with real-world settings, produced HRI knowledge with real-world applicability, and provides a compelling path towards specific follow-up studies. In addition to the statistical best practices it encourages [7], the VID2REAL HRI framework provides principled epistemological grounding for HRI research programs seeking to effectively and efficiently generate and build upon knowledge of complex sites, phenomena, and robotic systems.

II. RELATED WORK

Amidst longstanding calls for standard metrics and heightened scientific rigor in the field of HRI [8], [9], an increasing number of researchers are conducting methodologically oriented studies, seeking to better understand and improve the scientific knowledge generated from HRI research practices. These have included studies of reproducibility [10], the relative validity of lab-based and online studies [11], and the effects of different stimulus media, such as static images, moving images, and videos upon research results [12].

A range of frameworks have been advanced to guide research in ever-more situated and specialized contexts [13], [14]. Kunold [15], for instance, notes a lack of principled ways for turning research insights into design principles as the rationale for a framework for studying communication with social robots. Frameworks have been advanced to guide research on complex topics like group dynamics [16] or the formation of expectations of social robots [17].

HRI with mobile autonomous robots in real-world settings faces a range of methodological difficulties arising from the high dimensionality of potential factors. Factors found to have influenced real-world HRI with autonomous robots range from the robot characteristics [18] to robot physical behavior [19] to human-robot communication [20]. While real-world

[1]School of Information, The University of Texas at Austin, {eah13, ycchan}@utexas.edu
[2]Department of Computer Science, The University of Texas at Austin, {sadanandm, joydeepb, justinhart}@utexas.edu

HRI knowledge continues to advance, isolation and control of variables and context remains a substantial challenge, posing risks to the applicability of HRI research.

A potential solution to this problem is to use more highly controlled experimental setups in the laboratory or online [21]. The established validity of online studies for specific categories of HRI research questions [11], [12] is mitigated by its lack of fidelity of experience: seeing a video of an autonomous robot will rarely involve the full range of experiences present in real-world encounters. Laboratory studies, on the other hand, gain the ability to control variables and present in-person robotic stimuli to participants, but must trade the ecological validity of study setting inherent to research. Online studies have high statistical power and can be relatively well-controlled between participants, but have low ecological validity for studies of autonomous mobile robots specifically. Lab-based studies have the complex interactional possibilities of in-person studies, but trade statistical power to do so. Real-world studies' extremely high ecological validity is mitigated substantially. This situation, summarized in Table I, forces researchers to make difficult choices about the strengths and weaknesses of different study modalities.

Attempts to move outside the online/real-world dichotomy have yielded a range of partial solutions, including realistic laboratory settings, ad hoc incorporation of various modalities in a single study, or estimates of real-world validity based on published studies of very different design. While these approaches may increase real-world applicability, they can be difficult to evaluate and compare. HRI-specific study design frameworks responsive to the general move towards standardization [22] in the field are needed to improve the efficiency and validity of real-world HRI studies of autonomous robots.

III. Vid2Real HRI Framework

The Vid2Real HRI framework offers HRI researchers a principled way to use the complementary strengths of video-based and real-world research modalities in studies such as Human-Robot Interaction or incidental Human-Robot Encounters with autonomous robots. It characterizes the methodological particularities of each modality for HRI studies of autonomous robots in real-world settings; and defines the epistemological relationships between them. This section describes the framework's rationale, design, and potential and recommended uses.

A. Statistical and Epistemological Rationale

Statistical best practice for performing a study requires a power analysis to inform the estimated sample size required to achieve the desired statistical power [8], [23]. Past HRI research has utilized this practice, which becomes much easier if similar instruments are used for most studies. Wasted effort due to insignificant results is potentially avoidable through statistical techniques: accurate effect size estimates and power calculations can help inform researchers whether their design is likely to observe the intended effect. Statistical discipline, as urged by leading HRI methods reference sources [24] dramatically increases the efficiency and quality of research activity.

However, statistical best practice isn't enough. Without a principled framework for when and how to conduct a study online or in the field, researchers risk a non-optimal blend of these modalities' tradeoffs, shown in table I. In contrast, the Vid2Real HRI framework is designed to enhance real-world study outcomes by aligning the hypothesis and design of a video-based study with untested assumptions of a target real-world study. This alignment, described below and exemplified in section IV, is designed to achieve *commensurable* results, wherein results from each study are informative and predictive of the other. Vid2Real HRI thus takes an epistemological approach to the design of research studies, focusing on producing knowledge of unknowns and testing assumptions. In addition to facilitating statistical rigor, the framework enables a research team to efficiently and iteratively improve their knowledge of their subject and of their experimental intervention over a series of studies, within or across research publications. The Vid2Real HRI framework's focus on alignment helps to identify and realize the synergistic value of commensurable results in single studies and within a broader research program. Video-based studies are used to produce highly specific knowledge that help explain and isolate phenomena of interest in real-world HRI studies. Unexpected results or events in real-world studies generate hypotheses for refinement in future online studies. The Vid2Real HRI framework thus helps researchers use their current state of knowledge to determine which modality is most informative at a given stage of inquiry.

B. Framework Description

Figure 1 provides an overview of the Vid2Real HRI framework. We specify that video-based study designs should be aligned with a specific real-world study's hypotheses, design, and scenario, regardless of whether such a real-world study is actually performed. Deliberate alignment with a real-world study design makes video-based study results commensurable with and informative about the real-world settings of interest.

We define *commensurable* as a relative property that obtains between video and real-world studies that produce directly comparable and mutually predictive results. The framework aims to acheive commensurability between video-based and real-world study designs by aligning variable(s), scenario, and settings. This alignment ensures that similar results are obtainable, in principle, in real-world settings. The degree of commensurability realized between two studies aligned in this way is primarily a function of the fidelity with which the video-based encounter surrogate evokes the real-world encounter of interest for study participants.

We define *informative* as a property of the hypothesis tested in the video-based study, whereby the results have significant implications for real-world study design. The sense of information invoked here is from an information-theoretic sense: novel information is gained. Informativeness is, roughly, the expected surprise weighted by expected frequency. Informative study designs should thus test both assumptions, which are rarely expected to be violated but produce large amounts of novel information when they are, and empirical measurements, which produce small amounts of novel information reliably.

TABLE I: Summary of tradeoffs between video-based study and real-world study.

	Video-based Online Study	**Real-world Study**
PROS	• Reproducibility of stimulus • Fast recruitment • Ease of isolating factors **Higher Statistical Power**	• Multimodal data (e.g. sensors, observation) • Longitudinal potential • Site-specific context **Higher Ecological validity**
CONS	• Skills for creating video encounters required • Potential disconnection from specific contexts • Homogenous data (e.g. no sensors, observation) **Lower Ecological Validity**	• Inherently Between subjects • Labor-intensive • Difficult to isolate factors **Lower Statistical Power**
EXAMPLE STUDIES	Babel *et al.* [11], Randall and Sabanovic [12]	Liang *et al.* [18], Chen *et al.* [20]

The *hypothetical real-world study design* is defined as a version of the video-based study able to be conducted in person using the same variable(s), scenario, and setting. This study design is not carried out in all cases, would be carried out after the online results are obtained, and likely would be revised if carried out. The hypothetical real-world study design's primary function in the framework is as an alignment target for the video-based design, enabling the accomplishment of commensurability and informativeness. Follow-up real-world studies, if any, should have designs informed by the VID2REAL HRI study results (otherwise, the real-world study's value would be primarily methodological, as they are in the present research).

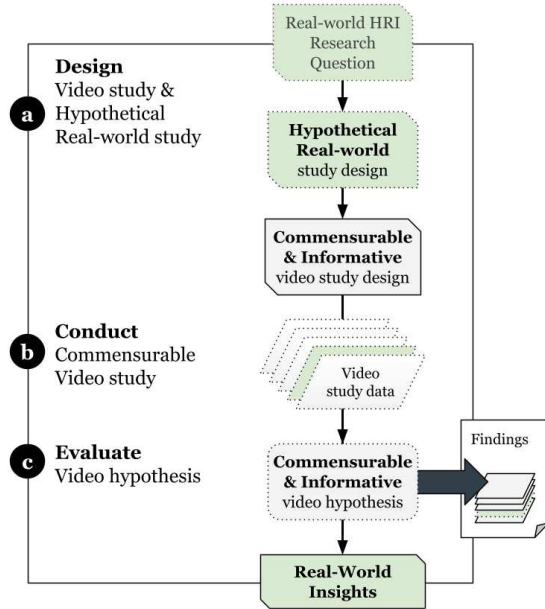The following paragraphs reference and explain the VID2REAL HRI phase labels shown in fig. 1.

Fig. 1: A schematic of the VID2REAL HRI process

a. Design. Beginning with an HRI research question, researchers identify a *hypothesis* and operationalize it in a *hypothetical real-world study design* (Figure 1a). Researchers then develop a *commensurable* video study hypothesis and study design, iteratively revising the designs as needed. The

video stimulus is conceived as an *encounter surrogate*, and can be judged by its capability to replicated participant experience intended in the hypothetical real-world study. The characteristics of videos may vary significantly by research question, so the VID2REAL HRI framework does not specify standard methods. An example of creating videos for human-robot encounters research is provided in Section IV.

b. Conduct and c. Evaluate. Once a commensurable video hypothesis and study design are ready and tested, the study can be conducted (Figure 1b). Note that standard best practices, such as obtaining institutional review board (IRB) approval and piloting the design should be followed, and are not shown in Figure 1. Due to the speed and relative cost-efficiency of participant recruitment, a large amount of data can be collected in a short amount of time. With this data collected, researchers can evaluate the commensurable video study hypothesis (Figure 1c), producing findings (these may be publishable or contribute to future publications, depending on their significance).

The findings of VID2REAL HRI study can of course be published without real-world replication if they are significant and novel. The insights produced by the alignment of study design with a specific real-world setting aid in interpretation of findings, the selection of future research questions, and design of real-world of video-based studies of the same or related real-world settings and scenarios. When warranted, the hypothetical real-world study design may be modified in light of the online video study results, supplementing the original results or constituting an independent research contribution. Conducting the hypothetical real-world study design with little modification, as done in the present paper, can characterize the methodological validity of the framework in specific applications where the applicability of the framework's validity (demonstrated below) is uncertain.

IV. VID2REAL HRI ONLINE STUDY

The VID2REAL HRI framework was applied to a specific HRE research study on the effects of socially compliant robot behavior conditions, operationalized as mixtures of verbal cues and body language, on pedestrian perceptions of the robot's social intelligence and willingness to open a door for the robot. The application study's illustrative value and relevance to the

strengths of the VID2REAL HRI framework is the focus of its presentation in this paper; full analysis of results and their implications for HRE research is presented elsewhere.

A. Application Online Study Design

The study scenario is an incidental encounter with a quadruped robot requiring human assistance to enter a building. Specifically, the robot is assisted by a human pushing a button to automatically open the doors. Since the nearby human has no obligation to assist the robot, we hypothesized that increased perceived social intelligence would increase human compliance with the robot's need for assistance.

Four robot behavioral conditions operationalized differentially socially compliant behaviors using verbal cues and body language:

- Baseline: The walking robot stops at the entrance and waits for the nearby human to push the button, without offering any indications that it needs assistance.
- Verbal: The walking robot stops at the entrance and says, *"Excuse me, can you open the door for me?"* to the nearby human, then waits for the human to push the button.
- Body Language: The walking robot stops at the entrance and turns its head facing the nearby human, then turns its head back to look at the door, then waits for the human to push the button.
- Body Language + Verbal: The walking robot stops at the entrance and turns its head facing the nearby human and says, *"Excuse me, can you open the door for me?"*, then turns its head back to look at the door and waits for the human to push the button.
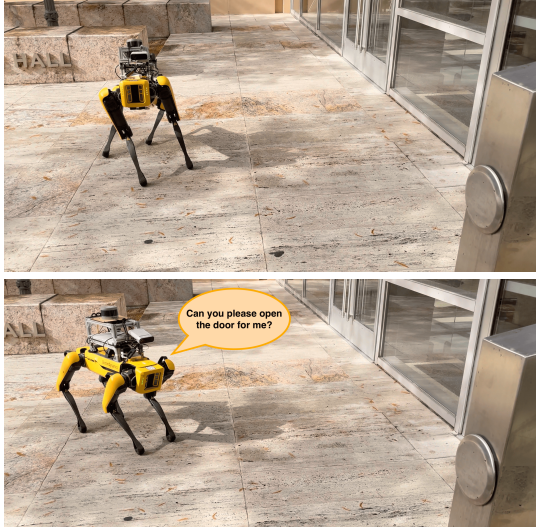

Fig. 2: Sample frames from the videos of the Body Language **(top)** and Verbal **(Bottom)** conditions.

1) Video and Real-World Alignment: To ensure the alignment of the video settings and the real-world scenarios, the videos are shot from a first-person point-of-view. This decision is mainly to support the need for the participants to imagine that they are the person encountering the robot. The robot motion is also reproduced across the conditions in order to maintain consistency. Additionally, the location where the videos are shot is the same as the place where the future in vivo HRE research will be conducted. Therefore, the ecological validity of this online study is reserved, and the results are commensurable. The videos and a detailed description of the study design are publicly available [25].

2) Questionnaire: The Perceived Social Intelligence (PSI) scale [26] was adopted as the measurement of the effects of robot body language and verbal cues. According to the authors' suggestions, research can adopt a subset of PSI that suits the research context and goals. We selected 8 of PSI's 13 Social Information Processing scales: (1) Social Competence, (2) Identifies Humans, (3) Identifies Individuals, (4) Recognizes Human Behaviors, (5) Adapts to Human Behaviors, (6) Predicts Human Behaviors, (7) Recognizes Human Cognitions, and (8) Adapts to Human Cognitions. Other scales were not chosen because they were not applicable to a first-person point-of-view video-based study.

In addition to PSI, two exploratory questions were presented with each study condition. The first was a five-point Likert question, "The person in the video represents how you would have responded in the situation", which was included to assess the participants' perceptions of the fidelity of the human's actions in the video to their own. The second was a free-form text response question that asked the participants to answer how they would have behaved in the situation that they just watched.

3) Study procedure: The participants were recruited on Prolific. They had to have no visual and hearing difficulties, a high level of English fluency, at least a 95% approval rate on Prolific, and currently live in the United States. The online study used a within-subject design, and the order of the study conditions was randomized to avoid the learning effect. The participants had to first consent to participate, and then they would be presented with one study video and the questionnaire. This step was randomly repeated four times. The total study lasted around 10 minutes. The University of Texas at Austin IRB reviewed and approved this study.

B. Application Study Results

1) Participants: The dataset of the application online study is publicly available [25] for review and reuse. The Prolific platform recruited 420 participants. However, upon careful investigation of response quality, 35 participants were excluded due to incomplete answers or cheating behaviors, leaving 385 valid responses for result analyses. Among the 385 participants, 194 participants identified as female, and 191 as male, the age ranged from 19 to 75 ($M = 38.53$, $SD = 12.86$).

2) Human Compliance: The human compliance was measured using the 5-point Likert exploratory question reported above. The average score are as follows: Baseline $M = 3.371$, $SD = 1.35$, Verbal: $M = 4.283$, $SD = 1.038$, Body Language $M = 3.698$, $SD = 1.281$, Body Language + Verbal $M = 4.376$, $SD = 0.992$. Using the one-way repeated measures ANOVA test with Bonferroni correction (corrected significance level is 0.008333)), the results showed that there was a significant main effect ($F(3, 1552) = 129.88$, $p < 0.001$, $\eta_p^2 = 0.252$). Pairwise comparisons were also used as ad-hoc analyses. The results showed that only the pair of Body Language + Verbal

condition and Verbal condition did not have a significant difference ($p = 0.03497$). Based on these findings, it was clear that adding body language and verbal commands were efficient ways of increasing human compliance in an HRE context. Meanwhile, the verbal command had a significantly stronger effect than the body language.

3) PSI ratings: Given the within-subject design, the one-way repeated measures ANOVA test with Bonferroni correction was used to evaluate the differences. The Body Language + Verbal condition was rated the highest ($M = 3.826$, $SD = 0.661$), followed by the Verbal condition ($M = 3.611$, $SD = 0.657$), followed by the Body Language condition ($M = 3.601$, $SD = 0.595$), and the Baseline condition was rated the lowest ($M = 3.151$, $SD = 0.784$). The ANOVA test showed a main effect on PSI overall rating ($F(3, 1552) = 144.26$, $p < 0.001$, $\eta_p^2 = 0.272$), indicating the robot's socially compliant behaviors had positive influences on its perceived social intelligence. Additionally, higher PSI seemed to indicate higher human compliance. Therefore, the higher socially intelligent a robot is, the more likely humans would comply and offer help when the robot asks for one. Please note that full analyses, such as pairwise comparisons and ANOVA tests on the subscales, are omitted from the scope of this study.

4) Power analyses: An initial contribution of the VID2REAL HRI framework is a reliable estimation of the number of participants needed to observe a significant effect. Real-world studies of incidental human-robot encounters almost always require between-subject designs due to the nature of the scenario. Between-subject designs typically require a larger sample size than within-subject designs, threatening researchers' ability to achieve their research goals. Drawing from the average of the 8 PSI scale ratings, power analyses of the Baseline condition compared to other conditions, and the Body Language condition compared to the Body Language + Verbal condition were conducted. The results are summarized in Table II. The results showed that 22 people would be required to reach a power of 0.95 in a between-subject study of Baseline and Body Language + Verbal condition. We also conducted an analysis of the Body Language condition compared to the Verbal condition, and the result was that 6902 people ($F(1, 384) = 0.723$, $p < 0.395$, $\eta_p^2 = 0.00188$) would be required. Please note that a between-subject design is chosen because it can better reflect the natural settings of real-world scenarios than a within-subject design.

V. REAL-WORLD STUDY

A. Study Design

Following the VID2REAL HRI framework and the knowledge obtained from the online video study, the Baseline and Body Language + Verbal conditions with a between-subject design were chosen for the real-world study due to its achievable sample size. The real-world study took place at the exact same location where the videos were shot. The participants were instructed to start at the same location as in the video and walk in the same direction. After encountering the robot, the participants had the freedom to do whatever they felt natural as pedestrians who incidentally encountered the robot. After this interaction, the participants had to answer a questionnaire, which contained the same items, except for the two exploratory questions, which were excluded as they were not applicable. In replacement of the exploratory questions, a camera was set up to record the interaction session, specifically focused on how the participants reacted to the robot (*i.e.*, whether or not they complied).

The robot behaviors were also the same as in the video studies. However, given the challenges of achieving full robot autonomy in such a real-world scenario, and more importantly to ensure that the robot behaviour is as close as possible to the video, the robot was teleoperated using a Wizard-of-Oz method. One author of the present study was designated as the study manager, who faced and gave instructions to the participants and pretended to start the study session using a laptop that was visible to them, while another author of this paper teleoperated the robot staying at a place where they could perfectly view the study process while being hidden from the participants. The process and setup were validated with pilot tests. The real-world study was approved by The University of Texas at Austin IRB. Upon successful completion of the study, all participants were entered into a draw to stand a chance to win a $100 gift card, with one of them finally receiving the gift card.

B. Study Result

A total of 26 participants were recruited, 13 experienced the Baseline condition and 13 had the Body Language + Verbal condition, 13 identified as males, 13 as females. The age ranged from 18 to 28 ($M = 21.04$, $SD = 3.01$). Given the sample size and the between-subject design, the Mann-Whitney U test was used to compare the PSI ratings between the conditions (Baseline: $M = 3.433$, $SD = 0.680$, Body Language + Verbal: $M = 3.952$, $SD = 0.630$). As expected, the results showed a significant difference ($z = 1.9801$, $p = 0.047$, $r = 0.39$). In terms of human compliance, 11 out of 13 participants opened the door in the Body Language + Verbal condition, whereas only one participant did so in the Baseline condition. Using Fisher's exact test, the Body Language + Verbal condition had a significantly higher proportion of participants who opened the door ($p < 0.001$).

VI. DISCUSSION

A. VID2REAL HRI *Benefits in the Application Study*

The benefits of the VID2REAL HRI's Online Video modality on the application HRE study are demonstrated through the data results and the accumulated knowledge gained in the study material preparation process.

1) Aligned results: The video study's statistical results serve as both independent findings and empirical foundations for future in vivo research of autonomous robot HRI. The VID2REAL HRI process of designing a video-based study commensurable with a target real-world study shown in fig. 1 enabled this advantage by keeping real-world application a central concern in video study design. Aligning an online video study with a specific real-world study design is an effective way to disentangle the high-dimensionality nature of HRI research of autonomous robots in real-world settings. Future HRE research may also build upon the findings obtained from

TABLE II: Power analyses and the estimated n

Intervention	Control	n for $\beta = .95$	η_p^2	Needed increase of n
Body language + Verbal	Baseline	22	0.413	0.00%
Verbal	Baseline	32	0.3043	45.45%
Body language	Baseline	36	0.2782	63.64%
Body language + Verbal	Body language	70	0.164	218.18%
Body language	Verbal	6902	0.0019	31,272.73%

online video studies and expand its scope to examine other factors underlying social environments.

The results of the PSI scale comparisons and the power analyses are sources for condition selection, especially for the between-subject design nature of research in real-world settings. For instance, in a research context the same as the application study, we learn that a study able to use the Baseline and Body Language + Verbal conditions results for effect size estimation leads to a reasonable sample size ($n = 22$). This could inform future real-world studies or enable researchers to confidently select a multi-condition *NxM* design without losing statistical significance by dividing their online study population too finely.

Small effect sizes or non-significance between conditions revealed by our approach were also informative. A comparison between the effect of body language and verbal cues on pedestrian perceptions would be drastically informed by the small effect size, the *p* value, and the extremely large sample size ($n = 6,902$) implied by these results. This suggests that a new study design would be required to distinguish the effects of body language and verbal forms of socially compliant behavior upon human compliance. A VID2REAL HRI study could rapidly inform selection of behavioral conditions suitable for investigating such a question.

B. Optimizing Modality Strengths

VID2REAL HRI helped us navigate the strengths and weaknesses of video and real-world study modalities (in table I). This section surveys our approach and highlights insights for researchers seeking to adopt the VID2REAL HRI framework.

Preparing the video encounter surrogate in situ helped us refine the study design and proactively control the eventual real-world encounters. Our first videos revealed that pedestrians might not realize that the robot is attempting to have their attention, the door might be opened from inside, pedestrians might walk in front of the robot or follow it without interacting with it (in our case, offering help). During the real-world study, several of these issues happened again, making maintaining identical behavior across all study sessions difficult. These led to small variations to the location and timing of the robot's behavior. Participants' pace also naturally varied as well, causing the robot to miss the opportune moment to perform its behaviors. Some participants didn't see the body language, instead only hearing the verbal command. These exigencies are components of real-world studies' ecological validity, but could obscure valid findings if they obscure the effects of

the study conditions. Confirming the effect using online study before entering the field helped us anticipate and mitigate these potential confounds through study design, while the power analysis helped us scope the study with confidence. The seemingly 'methodologically better' alternative, starting with a real-world study, would have risked selecting conditions requiring infeasible sample sizes, resorting to a between-subjects design (diminishing ecological validity), and/or recruiting too few participants.

We view our online study videos as *encounter surrogates*, aiming to replicated the conditions of incidental encounters with high fidelity. We thus used a first-person perspective of the encounter, rather than a third-person view of someone encountering a robot. This removed the variable of the appearance of the human as a potential confound, and the possibility of the participants trying to interpret the human actor's behaviors and be biased. Additionally, the questionnaire instructed participants to imagine that they were the person encountering the robot.

C. Summary of Framework Benefit

The VID2REAL HRI research framework offered a principled way to design a video-based study of encounters with a quadruped robot aligned to a specific real-world setting. The online study's alignment with a specific real-world study design enabled it to efficiently produce valid knowledge of real-world encounters in a specific site. The online study results produced power analysis-based estimates of required sample size that accurately represented the statistical power obtained in the real-world study. Finally, the researchers' experiential insight into the importance of the direct 'gaze' of the robot for the effectiveness of body language cues as perceived by humans suggests limitations of the chosen platform that can be directly targeted in future research. Together, these improvements increased the validity of application study results to a specific real-world setting and clarified the ways the results can directly inform future work.

D. Limitations and Future Work

While our real-world results suggest that our encounter surrogate was sufficient to predict real-world encounter data, future research is needed to understand effects of video different techniques upon surrogates' fidelity. This study only used video-based encounter surrogates due to their simplicity and suitability for HRE research. Video-based surrogates will likely be less informative to studies involving extensive user

interaction with robots. Interactive study modalities such as simulation and VR/AR experiences may be effective encounter surrogates, but were not evaluated in this study.

VII. Conclusion

VID2REAL HRI is a research framework that leverages the complementary benefits of video-based and real-world autonomous robot HRI study designs. The VID2REAL HRI framework was used to design an online, video-based study aligned with a real-world study design in a specific site. The online study videos, conceived of as *encounter surrogates*, utilized a first-person video of autonomous behavior conditions planned for the real-world study. This validated key study assumptions about the body language and verbal conditions' effectiveness, provided valuable characterization of the chosen survey scales over potential alternatives, and enabled targeted refinement of the autonomy. The aligned online and real-world study designs produced commensurable and informative findings, as demonstrated by similarity of results and accuracy of sample size estimates. The framework's benefits compared favorably to potential alternatives, such as lab-based studies or longitudinal in-situ instrumentation of real-world sites. The VID2REAL HRI framework offers HRI researchers a principled way to use the complementary strengths of video-based and real-world research modalities to understand and improve real-world encounters with autonomous robots.

References

[1] E. Hauser, Y.-C. Chan, P. Chonkar, G. Hemkumar, H. Wang, D. Dua, S. Gupta, E. M. Enriquez, T. Kao, J. Hart, R. Mirsky, J. Biswas, J. Jiao, and P. Stone, ""what's that robot doing here?": Perceptions of incidental encounters with autonomous quadruped robots," in *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. New York, NY, USA: ACM, July 2023.

[2] L. Dole and W. Ju, "Face and ecological validity in simulations: Lessons from Search-and-Rescue HRI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, no. Paper 451. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–8.

[3] A. Biswas, A. Wang, G. Silvera, A. Steinfeld, and H. Admoni, "SocNavBench: A grounded simulation testing framework for evaluating social navigation." [Online]. Available: https://doi.org/10.1145/3476413

[4] A. M. H. Abrams, P. S. C. Dautzenberg, C. Jakobowsky, S. Ladwig, and A. M. Rosenthal-von der Pütten, "A theoretical and empirical reflection on technology acceptance models for autonomous delivery robots," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21. Association for Computing Machinery, pp. 272–280. [Online]. Available: https://doi.org/10.1145/3434073.3444662

[5] S. Modak, N. Patton, I. Dillig, and J. Biswas, "Synapse: Learning preferential concepts from visual demonstrations," *arXiv preprint arXiv:2403.16689*, 2024.

[6] Z. Hu, F. Lucchetti, C. Schlesinger, Y. Saxena, A. Freeman, S. Modak, A. Guha, and J. Biswas, "Deploying and evaluating llms to program service mobile robots," *IEEE Robotics and Automation Letters*, 2024.

[7] H. Kang, "Sample size determination and power analysis using the G*Power software," *J. Educ. Eval. Health Prof.*, vol. 18, p. 17, July 2021.

[8] C. L. Bethel and R. R. Murphy, "Review of human studies methods in HRI and recommendations," vol. 4, pp. 347–359. [Online]. Available: https://www.infona.pl/resource/bwmeta1.element. springer-6b23bd21-1cd1-3179-b67e-ad4cda63b545

[9] S. N. Woods, M. L. Walters, K. L. Koay, and K. Dautenhahn, "Methodological issues in HRI: A comparison of live and video-based methods in robot to human approach direction trials," in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 51–58. [Online]. Available: http://dx.doi.org/10.1109/ROMAN.2006.314394

[10] H. Gunes, F. Broz, C. S. Crawford, A. R.-v. der Pütten, M. Strait, and L. Riek, "Reproducibility in human-robot interaction: Furthering the science of HRI," vol. 3, pp. 281–292. [Online]. Available: http://dx.doi.org/10.1007/s43154-022-00094-5

[11] F. Babel, J. Kraus, P. Hock, H. Asenbauer, and M. Baumann, "Investigating the validity of online robot evaluations: Comparison of findings from an one-sample online and laboratory study," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21 Companion. Association for Computing Machinery, pp. 116–120. [Online]. Available: https://doi.org/10.1145/3434074.3447141

[12] N. Randall and S. Sabanovic, "A picture might be worth a thousand words, but it's not always enough to evaluate robots," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. Association for Computing Machinery, pp. 437–445. [Online]. Available: https://doi.org/10.1145/3568162.3576970

[13] A. Rosenthal-von der Pütten and A. M. H. Abrams, *Empirical Methods in the Social Science for Researching Socially Interactive Agents*, 1st ed. Association for Computing Machinery, vol. 37, pp. 19–76. [Online]. Available: https://doi.org/10.1145/3477322.3477325

[14] M. F. Damholdt, C. Vestergaard, and J. Seibt, "Testing for 'anthropomorphization': A case for mixed methods in human-robot interaction," in *Springer Series on Bio- and Neurosystems*. Cham: Springer International Publishing, 2020, pp. 203–227.

[15] L. Kunold, "Seeing is not feeling the touch from a robot," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. ieeexplore.ieee.org, pp. 1562–1569. [Online]. Available: http://dx.doi.org/10.1109/RO-MAN53752.2022.9900788

[16] A. M. H. Abrams and A. M. R.-v. der Pütten, "I–C–E framework: Concepts for group dynamics research in human-robot interaction," vol. 12, pp. 1213–1229. [Online]. Available: https://doi.org/10.1007/s12369-020-00642-z

[17] J. M. Berzuk and J. E. Young, "Clarifying social robot expectation discrepancy: Developing a framework for understanding how users form expectations of social robots," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. Association for Computing Machinery, pp. 231–233. [Online]. Available: https://doi.org/10.1145/3568294.3580078

[18] C. Liang, A. E. Ricci, H. Kress-Gazit, and M. F. Jung, "Lessons from a robot asking for directions in-the-wild," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. Association for Computing Machinery, pp. 617–620. [Online]. Available: https://doi.org/10.1145/3568294.3580159

[19] Y. Okafuji, Y. Ozaki, J. Baba, J. Nakanishi, K. Ogawa, Y. Yoshikawa, and H. Ishiguro, "Behavioral assessment of a humanoid robot when attracting pedestrians in a mall," *Adv. Robot.*, vol. 14, no. 7, pp. 1731–1747, July 2022.

[20] Z. Chen, T. Fan, X. Zhao, J. Liang, C. Shen, H. Chen, D. Manocha, J. Pan, and W. Zhang, "Autonomous social distancing in urban environments using a quadruped robot," *IEEE Access*, vol. 9, pp. 8392–8403, 2021.

[21] B. Holman, A. Anwar, A. Singh, M. Tec, J. Hart, and P. Stone, "Watch where you're going! gaze and head orientation as predictors for social robot navigation." IEEE, 2021, pp. 3553–3559.

[22] M. Schrum, M. Ghuy, E. Hedlund-botti, M. Natarajan, M. Johnson, and M. Gombolay, "Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices," vol. 12, pp. 1–32. [Online]. Available: https://doi.org/10.1145/3572784

[23] A. Kahng, M. K. Lee, R. Noothigattu, A. Procaccia, and C.-A. Psomas, "Statistical foundations of virtual democracy," in *Proceedings of Machine Learning Research*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, pp. 3173–3182. [Online]. Available: http://proceedings.mlr.press/v97/kahng19a.html

[24] C. L. Bethel, Z. Henkel, and K. Baugus, *Conducting Studies in Human-Robot Interaction*. Springer International Publishing, pp. 91–124. [Online]. Available: https://doi.org/10.1007/978-3-030-42307-0_4

[25] Y.-C. Chan, S. Modak, E. Hauser, J. Biswas, and J. Hart, "Community embedded robotics: Vid2Real an online video dataset about perceived social intelligence in human robot encounters," 2024.

[26] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, A. Fink-Armold, S. B. Banisetty, and D. Feil-Seifer, "Measuring the perceived social intelligence of robots," *J. Hum.-Robot Interact.*, vol. 9, no. 4, pp. 1–29, Sept. 2020.