2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Extraction of Latent Variables for Modeling Subjective Quality in Time-series Human-Robot Interaction

Yoshiaki Mizuchi[1], Taisuke Kobayashi[2] and Tetsunari Inamura[3]

Abstract— **This study presents a novel method for modeling subjective evaluation of the quality of interaction (QoI) by extracting explanatory variables that are not explicitly quantifiable by humans from human-robot behavior. The proposed method extracts latent variables that account for both explicit and tacit knowledge by performing maximum likelihood estimation to predict manually selected explanatory variables, alongside QoI score prediction, from time-series interaction data. In this study, we address three key questions: (i) whether the extraction of latent variables improves accuracy compared to conventional regression analysis, (ii) whether implicit variables, beyond those selected by humans, play a significant role, and (iii) whether human-selected explanatory variables are necessary in explaining subjective assessment scores. The results of comparisons across several learning conditions demonstrate that incorporating tacit knowledge variables, uncorrelated with traditional explanatory variables, enhances the accuracy of QoI estimation. This study contributes by enabling data-driven extraction of explanatory variables, revealing the influence of tacit knowledge on QoI estimation, and highlighting the importance of both top-down and bottom-up approaches in accurately estimating subjective evaluations of QoI.**

I. INTRODUCTION

Conveying information to others is essential for cooperative tasks, such as teaching work procedures, sharing roles, and requesting assistance when encountering errors. When engaging in a cooperative task with another person, we can intuitively sense whether things are progressing well. This ability to evaluate the quality of interaction (QoI) is crucial for interactive service robots that need to effectively cooperate with humans. Since QoI fundamentally depends on human impressions, subjective evaluation is indispensable for properly assessing QoI. Standard practice typically involves using questionnaires to conduct these evaluations. However, questionnaire-based subjective evaluation is time-consuming and impractical in specific scenarios, such as real-time decision-making based on another person's responses.

*This work was supported by JST [Moonshot R&D][Grant Number JPMJMS2034], a project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization(NEDO), ROIS NII Open Collaborative Research 2023-23FP05, and Kayamori Foundation of Informational Science Advancement

[1]Yoshiaki Mizuchi is with the College of Engineering, Tamagawa University, 6-1-1 Tamagawagakuen, Machida, Tokyo, 194-8610, Japan mizuchi@eng.tamagawa.ac.jp

[2]Taisuke Kobayashi is with the National Institute of Informatics (NII) and with The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan kobayashi@nii.ac.jp

[3]Tetsunari Inamura is with the Brain Science Institute, Tamagawa University, 6-1-1 Tamagawagakuen, Machida, Tokyo, 194-8610, Japan inamura@lab.tamagawa.ac.jp

To address this issue, we proposed a modeling approach for QoI evaluation that approximates human subjective assessment based on the behavior data [1]. In the previous study, through a VR-based competition task [2], [3], [4], [5] where a virtual robot guides human daily life actions using verbal instructions and gestures, we collected interaction behavior data and demonstrated the effectiveness of the proposed approach. However, the regression model used in the previous work relied on manually selecting explanatory variables, limiting the estimation of subjective scores to variables that can be explicitly quantified, such as time taken and distance traveled. Consequently, there remain concerns that variables selected by humans may not be optimal for explaining subjective assessment results, and essential implicit variables relevant to explaining evaluation values may be overlooked.

This study proposes a method for extracting explanatory variables from interaction behavior data that humans cannot explicitly quantify to achieve a more explanatory estimation of subjective QoI. Additionally, we address the following research questions: (i) whether extracting latent variables from behavior data improves accuracy compared to simple multiple regression analysis, (ii) whether there are implicit variables, beyond those selected by humans, that are significant for explaining subjective assessment scores, and (iii) whether explanatory variables selected by humans are essential for explaining subjective assessment scores.

II. RELATED WORK

Although various evaluation metrics for human-robot interaction have been proposed thus far [6], subjective scores are typically assessed using questionnaires. Mayima et al. [7] proposed an evaluation criterion for the quality of interaction based on behavior data. However, the proposed criterion does not account for human subjectivity, which cannot be directly observed. Previous studies, such as Kanda et al. [8] and our previous work [1] analyzed the relationship between subjective evaluation results and behavior data but focused solely on observable explanatory variables. In contrast, this study addresses explanatory variables that are not explicitly quantified by humans and aims to estimate subjective evaluation results for the quality of interaction based on behavior data.

Variational autoencoder (VAE) [9] is widely known as a deep learning technique for extracting latent variables hidden in data. As an extension of VAE, attribute regularization, which constructs a latent space explicitly considering the attributes of the data, has been proposed [10], [11]. Among

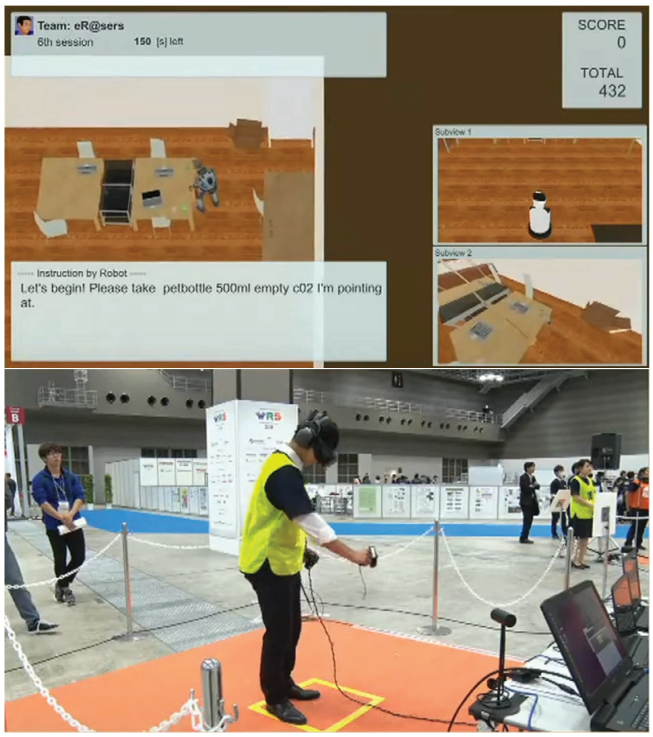


Fig. 1. Screenshots of the competition system and test subject.

them, methods in which some latent variables are intentionally made uncorrelated with the attributes [12], [13], [14] are closely related to our proposed method. However, these techniques target image data, and the loss functions they minimize are cumbersome. On the other hand, the proposed method is based on VAE for the analysis of time-series data [15] and utilizes a gradient reversal layer (GRL) [16] to achieve a simple implementation.

III. INTERACTION DATASET

A. Interaction Scenario

We focused on a competition task named *Human Navigation* [2], [3], [4], [5], which aims to evaluate a robot's ability to provide clear and user-friendly explanations to guide daily human actions. Figure 1 shows the competition system and a test subject. The test subject logs in to an avatar in the VR scene using a VR headset and interacts with the virtual environment and robot.

Figure 2 illustrates the task procedure of the Human Navigation task. The goal of the task is for the human user (i.e., test subject) to identify a target object and bring it to a specified destination. However, the test subject is not provided with information about room layouts, the target object, or the destination. At the beginning of each session, only the robot is informed about the positions and orientations of furniture, graspable objects, the target object, and the destination area. The test subject performs the task based on the instructions provided by the robot. If the robot's instructions are vague, the test subject may wander and fail to complete the task.

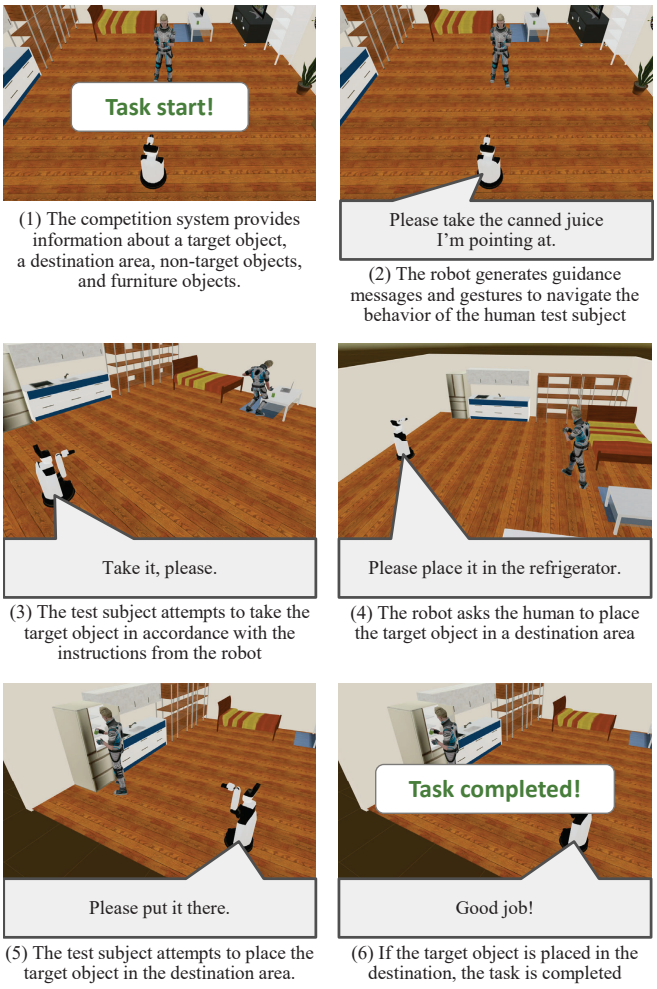


Fig. 2. Task procedure of the Human Navigation task.

B. Interaction Behavior Data

We collected various quality of interaction behavior data through three different approaches [17]. One part of the data was recorded during the World Robot Summits 2018 competition [3], where robot controllers generated instructions and gestures based on algorithms developed by each participating team. The quality of the interactions was relatively low, as developing effective instruction generators remains a challenge. Another part of the data was collected through unconstrained oral human-human interactions, where a human operator provided verbal instructions instead of a virtual robot. This approach resulted in high-quality interaction behavior data due to the flexibility of human instructions. The remaining data was collected through constrained human-human interactions using a graphical user interface (GUI)–based instruction generator. In this approach, human operators generated instructions by selecting template sentences and incorporating words such as object names, properties, positional relationships, directions, and distances. This approach, with its restricted instruction strategies and vocabulary, produced interaction behavior data of intermedi-

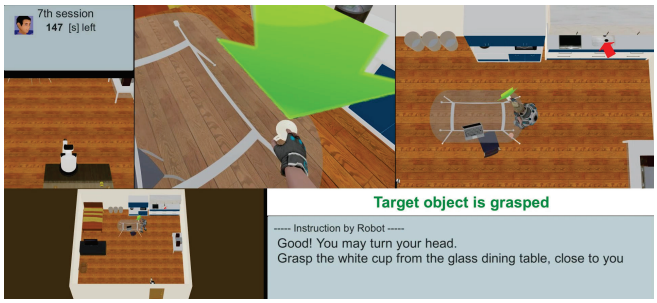


Fig. 3. Video for evaluation.

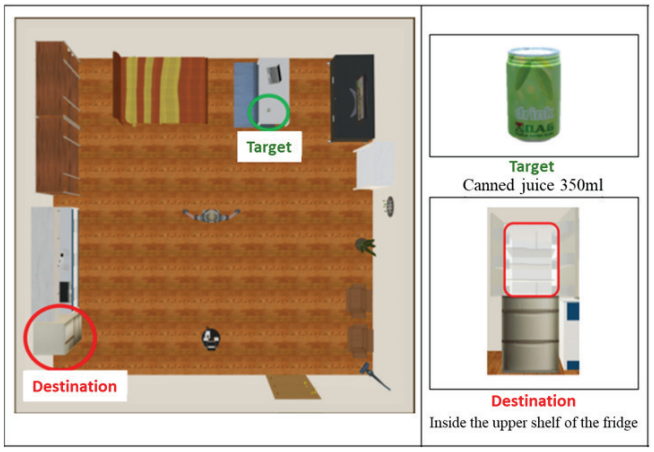
ate quality between the other two approaches.

The behavior data included the movements of the avatar and robot (i.e., the 6-DOF pose of each joint), 6-DOF poses of the graspable objects, speech sentences provided by the robot, and event information (e.g., the grasp of a target object).

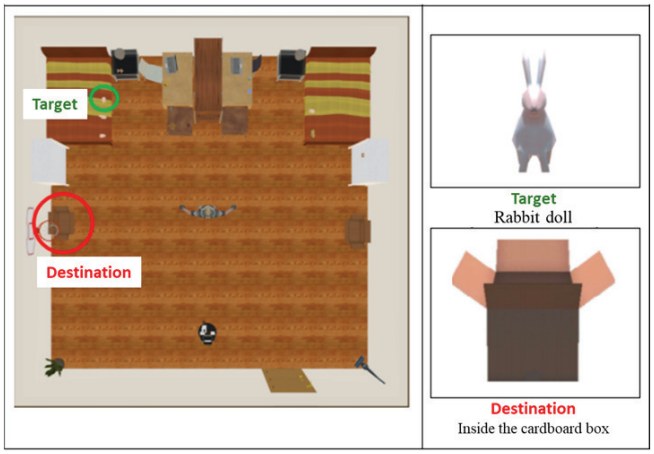
C. Subjective QoI Data

As shown in Fig. 3, we created playback videos of the collected interaction behaviors and allowed third parties to evaluate them [17]. We selected interaction behavior data performed in three room layouts. Figure 4 shows the selected room layouts. In layout A, the target object is placed on a table in front of the avatar. The test subject is required to open the upper door of the fridge to place the target object inside. In layout B, non-target objects of the same type as the target object are positioned near the target object. Additionally, a cardboard box that is not the intended destination is placed on the opposite side of the room. In layout C, a plastic bottle with a similar design but a differently colored label is placed near the target plastic bottle. Furthermore, another trash can of the same type as the intended destination is positioned next to the destination trash cans. Hence, robots need to provide information that enables the test subjects to identify the target object, as their performance depends on the instruction sentences provided by the robots. For each layout, sixteen sessions of interaction behavior were recorded under the same target object and destination conditions: six sessions were collected through human-robot interaction in a robot competition, six sessions through GUI-based human-human interaction, and four sessions through unrestricted oral human-human interaction. We asked 24 evaluators to review these videos and rate the overall quality of interaction for each session using a 7-point Likert scale, based on the question: "The quality of the overall human-robot interaction was high."

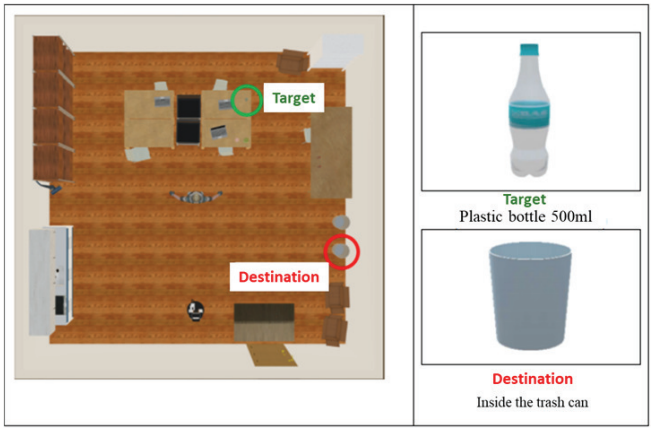
Each evaluator assessed the videos of the sixteen sessions recorded in a specific layout.



Layout A



Layout B



Layout C

Fig. 4. Room layouts.

IV. PROPOSED METHOD

A. Architecture

Figure 5 shows the proposed method for extracting explanatory variables and estimating subjective QoI scores. The time-series data of interaction behavior in a single session is denoted as $X = [x_1, x_2, \ldots, x_T]$ with T the length. We define the observed data concatenated over H steps as $\chi_\tau^H = \left[x_{H(\tau-1)+1}, x_{H(\tau-1)+2}, \cdots, x_{H\tau}\right]$, while augmenting the time-series data with x_T at the end $(T \bmod H)$ times. An encoder extracts the local latent variable z_τ hidden in χ_τ^H,

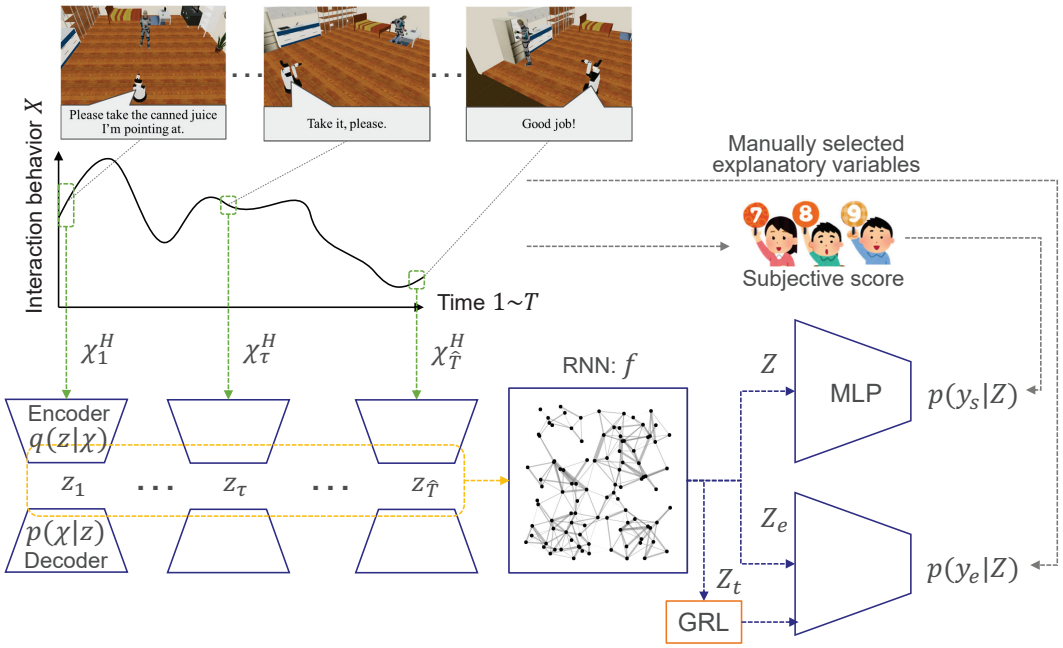


Fig. 5. Network architecture for extracting explanatory variables and predicting subjective QoI.

which can be reconstructed by passing z_τ through a decoder.

Additionally, the time-series data of the extracted latent variables $\left[z_1, \ldots, z_\tau, \ldots, z_{\hat{T}}\right]$ ($\hat{T} = \lceil T/H \rceil$) are fed into a Recurrent Neural Network (RNN) in order, approximating the global latent variable $Z \simeq \mathrm{RNN}\left(z_1, \ldots, z_\tau, \ldots, z_{\hat{T}}\right)$. Z is fed into a Multi-Layer Perceptron (MLP) to predict the subjective evaluation result y_s. Furthermore, Z is divided into two parts, $Z = [Z_e, Z_t]$: Z_e is for directly representing the manually designed explanatory variables y_e; while Z_t is for revealing the information uncorrelated with y_e (i.e., the tacit knowledge) by passing through a Gradient Reversal Layer (GRL) [16]. Z_e and Z_t are then used to predict y_e with another MLP.

B. Derivation of loss function

Based on the above information processing architecture, we formulate a loss function for learning all network modules (i.e., the encoder, decoder, RNN, and two MLPs) with parameters θ. Specifically, we consider X, y_e, and y_s as random variables and maximize their log-likelihoods, leading to a variational lower bound based on VAE [9], [15]. This negative value is the loss function to be minimized, as follows:

$$
\begin{aligned}
\ln p(X, y_e, y_s) &= \ln \mathbb{E}_{p(Z)}\left[p(X|Z)p(y_e, y_s|Z)\right] \\
&\geq \mathbb{E}_{q(Z|X)}\left[\ln p(X|Z) + \ln \frac{p(Z)}{q(Z|X)}\right] \\
&+ \mathbb{E}_{q(Z|X)}\left[\ln p(y_e|Z) + \ln p(y_s|Z)\right] \\
&= \sum_{\tau=1}^{\hat{T}} \mathbb{E}_{q(z_\tau|\chi_\tau^H)}\left[\ln p(\chi_\tau^H|z_\tau) + \ln \frac{p(z_\tau)}{q(z_\tau|\chi_\tau^H)}\right] \\
&+ \mathbb{E}_{q(Z|X)}[\ln p(y_e|Z_e, \mathrm{GRL}(Z_t)) + \ln p(y_s|Z)] \\
&= -\mathcal{L}(\theta) \qquad (1)
\end{aligned}
$$

where $q(z_\tau|\chi_\tau^H)$ is the encoder given as the variational posterior (with $p(z_\tau)$ the prior), $p(\chi_\tau^H|z_\tau)$ is the decoder, $p(y_e|Z_e, \mathrm{GRL}(Z_t))$ and $p(y_s|Z)$ are represented by the MLP modules. $q(Z|X)$ is given as the production of $q(z_\tau|\chi_\tau^H)$ with $\tau = 1, 2, \ldots, \hat{T}$, and Z can be sampled from it by that z_τ sampled from $q(z_\tau|\chi_\tau^H)$ ($\tau = 1, 2, \ldots, \hat{T}$) are passed to the RNN module. Note that we assume that a certain value of H guarantees enough independence to adequately reconstruct the observed data.

The loss function is computed by approximating the expectation operations with a Monte Carlo method using stochastic variables sampled from the corresponding probabilities. To minimize this loss, we optimize θ using a combination of a backpropagation with reparameterization trick [9] and a stochastic gradient descent. As a result, Z embeds the important information of X, while it is calibrated to represent y_s. In addition, Z_e and Z_t are further corrected to include and exclude the information of y_e, respectively.

V. EXPERIMENT

A. Dataset

The interaction behavior data X was represented as a 25-dimensional vector, including timestamps, the positions and orientations of the human's head and both hands, and the position and orientation of the target object. The data was formatted into a time-series with $25H$ dimensions, where H was set to 5. Additionally, the center position of the destination area where the target object is to be placed was included, resulting in a final input data dimension of 128 for χ. The following 10-dimensional explanatory variables, denoted as y_e, were manually selected: (1) whether the subject was able to grasp the target object within the time limit, (2) the time taken for the subject to grasp the target

object, (3) whether the subject completed the task within the time limit, (4) the time taken to complete the task, (5) the number of instructions issued by the robot, (6) the number of times the subject grasped an incorrect object, (7) the number of times the subject requested clarification, (8) the number of times the robot performed pointing gestures, (9) the cumulative change in the orientation of the subject's face, and (10) the cumulative distance traveled by the subject.

A total of 384 data were included in the dataset for this scenario: combinations of (a) three layouts, (b) three interaction methods, and (c) seven to nine evaluators. We selected a total of nine data with each layout and each interaction method as a validation dataset, a different total of nine data as a test dataset, and the remaining 366 data as a training dataset. X, y_e, and y_s were standardized by the sample mean and standard deviation of the training dataset, respectively.

B. Learning conditions

The activation function used was RMSNorm [18] + Squish [19]. The probability distribution models for all variables were set to diagonal normal distributions. To ensure that its scale parameter is positive, a Squareplus transformation [19] was applied. The encoder and decoder each consisted of three fully connected layers with 128, 64, 32 neurons. The latent dimension was $|Z| = 32$. As the RNN module to transform from $\left[z_1, \ldots, z_\tau, \ldots, z_{\hat{T}}\right]$ to Z was designed with a single-layer bidirectional LSTM with 64 units. For the MLP modules for predicting the subjective evaluation results and the manually selected explanatory variables, a single fully connected layer with 64 neurons were employed, respectively.

The weights for loss terms, the reconstruction, KL regularization, maximum likelihood estimation for y_e and y_s, were set to 1 except 0.1 for the KL regularization. The weight for the GRL [16] was set to 1. The AdaTerm algorithm [20] was used for stochastic gradient descent, with its default parameters (e.g. a learning rate $alpha = 10^{-3}$). A single data with (X, y_e, y_s) was sampled from the shuffled training dataset for each epoch, and training was conducted for 100 epochs without early stopping.

To evaluate the effectiveness of the proposed method, we compared the prediction accuracy of y_s across four conditions: MLP, TT, EE, and ET. In the MLP condition, y_s is directly predicted from y_e by the maximum likelihood estimation. Thus, this condition replaces the conventional multiple regression analysis with maximum likelihood estimation. In the TT condition, all variables in Z were regarded as tacit knowledge variables, that is, the maximum likelihood estimation for y_e was ignored. In the EE condition, all variables in Z were regarded as explicit knowledge variables, and the maximum likelihood estimation of y_e was performed, but the GRL was not applied to Z_t. In the ET condition, half of the variables in Z were regarded as explicit knowledge and the other half as tacit knowledge, as the proposed method.

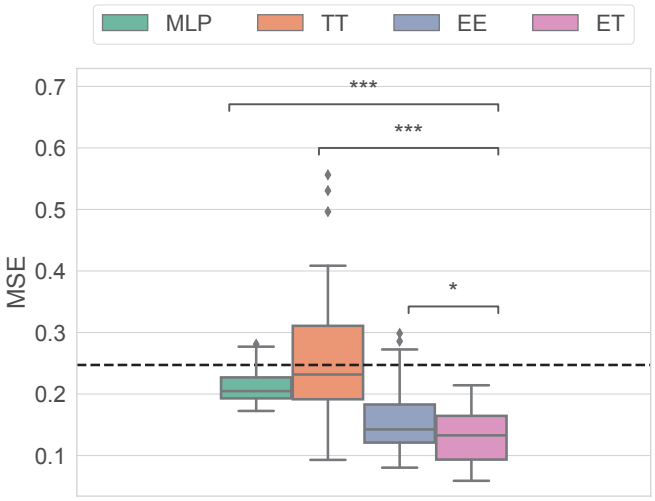


Fig. 6. Mean square errors (MSE) in predicting subjective QoI scores. The box plots show the distribution of MSE values for each condition. Asterisks above the boxes indicate significant differences between conditions as determined by pairwise comparisons using a one-sided Mann-Whitney U-test. Significance levels are denoted as follows: "***" as $p < 10^{-3}$, "**" as $p \in [10^{-3}, 10^{-2})$, "*" as $p \in [10^{-2}, 0.05)$, "+" as $p \in [0.05, 0.1)$, and "ns" as $p > 0.1$

C. Results

Figure 6 presents the boxplots of mean square errors (MSE) in predicting subjective QoI scores y_s for each condition with 36 initialization patterns by different random seeds. The dotted line is the baseline result of multiple linear regression based on manually selected explanatory variables.

The MSE under the ET condition was significantly lower than all other conditions at a 5% significance level. Under the TT condition, where manually selected explanatory variables were not considered, the MSE was higher than in all other conditions, including the simple multiple regression model. Therefore, it is demonstrated that simply extracting latent variables from behavior data is insufficient for accurate prediction, and that the selection of explanatory variables by humans, based on explicit knowledge, is crucial for modeling subjective QoI evaluation. Under the EE condition, where latent variables were extracted from manually selected explanatory variables, the MSE was lower than that of the MLP condition. This indicates that only manually selected explanatory variables cannot capture some elements, and that extracting latent variables has its merits. The MSE under the ET condition, where latent variables uncorrelated with the manually selected explanatory variables were extracted, was lower than that under the EE condition. In addition, the MSE for the explanatory variables of the ET condition is as small as that of the EE condition (i.e., less than 0.01). These results suggest that there are tacit knowledge variables, outside of those considered by humans (i.e., explicit knowledge), that influence the prediction of subjective QoI scores. Therefore, the effectiveness of the proposed method, which involves extracting explanatory variables that include tacit knowledge variables from behavior data in addition to manually selected explanatory variables, has been demonstrated.

VI. DISCUSSION

Evaluation criteria for approximating subjective QoI should be tailored to each specific domain. However, manually designing complex explanatory variables is a challenging and time-consuming task. A data-driven method for designing explanatory variables from behavior data, established in this study, greatly enhances the applicability of modeling subjective QoI evaluation, marking a significant contribution.

The proposed method facilitates the design of variables that were previously difficult for humans to explicitly quantify, influencing the estimation of subjective QoI. Nonetheless, this study has not yet identified which specific elements of interaction behavior correspond to each latent variable. Identifying these variables is a future challenge to help developers and robots understand which aspects of interaction behavior should be improved.

The proposed method can be improved for more estimation accuracy and explainability. While we used LSTM to transform sequential interaction behavior data into latent variables, using a Transformer model might increase the expressive power. Additionally, the objective variable was only the overall quality of interaction. The authors have identified the metrics that evaluators use when assessing overall interaction quality [21], and incorporating these metrics as objective variables may allow for more interpretable evaluation predictions.

VII. CONCLUSIONS

This paper proposed a method for modeling the subjective quality of interaction (QoI) by extracting latent variables from time-series interaction behavior data. The proposed method involved extracting explanatory variables, including those not considered by humans (i.e., tacit knowledge), to explain subjective QoI scores. The results of comparisons across several learning conditions demonstrated that incorporating tacit knowledge variables, which were uncorrelated with manually selected explanatory variables (i.e., explicit knowledge), improved the accuracy of QoI score estimation. The contributions of this study were threefold: (i) enabling the data-driven extraction of explanatory variables from time-series interaction behavior for modeling subjective QoI evaluation; (ii) revealing the existence of tacit knowledge variables that influenced QoI score estimation; and (iii) demonstrating that both top-down design of explanatory variables based on human implicit knowledge and bottom-up extraction of variables from behavior data were important for accurate QoI score estimation.

REFERENCES

[1] Y. Mizuchi and T. Inamura, "Optimization of criterion for objective evaluation of hri performance that approximates subjective evaluation: a case study in robot competition," *Advanced Robotics*, vol. 34, no. 3-4, pp. 142–156, 2020.

[2] T. Inamura and Y. Mizuchi, "Robot Competition to Evaluate Guidance Skill for General Users in VR Environment," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2019, pp. 552–553.

[3] H. Okada, T. Inamura, and K. Wada, "What competitions were conducted in the service categories of the world robot summit?" *Advanced Robotics*, vol. 33, no. 17, pp. 900–910, 2019.

[4] T. Inamura, Y. Mizuchi, and H. Yamada, "VR platform enabling crowdsourcing of embodied HRI experiments – case study of online robot competition," *Advanced Robotics*, vol. 35, no. 11, pp. 697–703, 2021.

[5] Y. Mizuchi, H. Yamada, and T. Inamura, "Evaluation of an online human-robot interaction competition platform based on virtual reality – case study in rcap2021," *Advanced Robotics*, vol. 37, no. 8, pp. 510–517, 2023.

[6] P. Damacharla, A. Y. Javaid, J. J. Gallimore, and V. K. Devabhaktuni, "Common metrics to benchmark human-machine teams (hmt): A review," *IEEE Access*, vol. 6, pp. 38 637–38 655, 2018.

[7] A. Mayima, A. Clodic, and R. Alami, "Towards robots able to measure in real-time the quality of interaction in HRI contexts," *International Journal of Social Robotics*, vol. 14, no. 3, pp. 713–731, 2022.

[8] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Development and evaluation of interactive humanoid robots," *Proc. of the IEEE*, vol. 92, no. 11, pp. 1839–1850, 2004.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[10] A. Creswell, Y. Mohamied, B. Sengupta, and A. A. Bharath, "Adversarial information factorization," *arXiv preprint arXiv:1711.05175*, 2017.

[11] G. Hadjeres, F. Nielsen, and F. Pachet, "Glsr-vae: Geodesic latent space regularization for variational autoencoder architectures," in *2017 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2017, pp. 1–7.

[12] J. Klys, J. Snell, and R. Zemel, "Learning latent subspaces in variational autoencoders," *Advances in neural information processing systems*, vol. 31, 2018.

[13] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, and Z. Tu, "Guided variational autoencoder for disentanglement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7920–7929.

[14] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "Diva: Domain invariant variational autoencoders," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 322–348.

[15] S. Tonekaboni, C.-L. Li, S. O. Arik, A. Goldenberg, and T. Pfister, "Decoupling local and global representations of time series," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8700–8714.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[17] Y. Mizuchi, K. Iwami, and T. Inamura, "VR and GUI based Human-Robot Interaction Behavior Collection for Modeling the Subjective Evaluation of the Interaction Quality," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 375–382.

[18] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] T. Kobayashi and T. Aotani, "Design of restricted normalizing flow towards arbitrary stochastic policy with computational efficiency," *Advanced Robotics*, vol. 37, no. 12, pp. 719–736, 2023.

[20] W. E. L. Ilboudo, T. Kobayashi, and T. Matsubara, "Adaterm: Adaptive t-distribution estimated robust moments for noise-robust stochastic gradient optimization," *Neurocomputing*, vol. 557, p. 126692, 2023.

[21] Y. Mizuchi, Y. Tanno, and T. Inamura, "Designing Evaluation Metrics for Quality of Human-Robot Interaction in Guiding Human Behavior," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, 2023, pp. 39–45.