2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

# Perceive, Represent, Generate: Translating Multimodal Information to Robotic Motion Trajectories

Fábio Vital*,[1] Miguel Vasco* Alberto Sardinha* and Francisco Melo*

***Abstract*— We present *Perceive-Represent-Generate* (PRG), a novel three-stage framework that maps perceptual information of different modalities (e.g., visual or sound), corresponding to a series of instructions, to a sequence of movements to be executed by a robot. In the first stage, we perceive and preprocess the given inputs, isolating individual commands from the complete instruction provided by a human user. In the second stage we encode the individual commands into a multimodal latent space, employing a deep generative model. Finally, in the third stage we convert the latent samples into individual trajectories and combine them into a single dynamic movement primitive, allowing its execution by a robotic manipulator. We evaluate our pipeline in the context of a novel robotic handwriting task, where the robot receives as input a word through different perceptual modalities (e.g., image, sound), and generates the corresponding motion trajectory to write it, creating coherent and high-quality handwritten words.**

## I. INTRODUCTION

Recent advancements in artificial perception [1] and actuation [2] have fostered the widespread use of robotic systems in various tasks, such as autonomous driving [3], industrial manufacturing [4], and medical [5] or education [6] scenarios. Furthermore, the number of tasks that require collaboration between robots and human users is expected to increase, raising significant challenges regarding the quality of their interaction and the mismatch between their perceptual, cognitive, and actuation capabilities. To improve the efficiency of robots in such scenarios, these systems can be provided with additional sensors supplying multimodal information of its environment [7]. The access to additional perceptual information is fundamental as humans often employ multiple communication channels in these scenarios, such as speech and non-verbal communication [8].

In this work, we address the problem of *how to translate multimodal commands* provided by a human user through different communication channels to a *movement* executed by a robotic agent. In particular, we consider a scenario where the human user provides high-dimensional perceptual data (e.g., sound, images) related to the task, such as the words in a handwriting task. The agent's role is to decompose the raw observations (e.g., the letter sequence forming a word) and generate the corresponding motion trajectory. Moreover, the performance of the agent must be robust to missing modality information, as the human user may not employ all possible communication channels during task execution.

To address such problem, we contribute a novel three-stage framework *Perceive-Represent-Generate* (PRG) that maps multimodal perceptual information provided by a human user to a corresponding motion trajectory executed by the robot. Initially, the agent *perceives* the environment, collecting and processing the raw multimodal observations into a sequence of individual task components (e.g., letters in a word). Subsequently, in the second stage, the agent *represents* the individual task components, mapping them into a multimodal latent space, encoded by a deep generative model. Finally, in the third stage, the agent *generates* and merges motion information decoded from the latent representations to execute the final motion.

We instantiate our PRG pipeline in a novel multimodal scenario (*Robotic Dictaphone*) where the robot is provided with textual information (through a combination of sound, image, or motion observations) and generates a single motion trajectory to write the target word, mimicking human handwriting. We perform quantitative and qualitative evaluations of PRG in the *Robotic Dictaphone* scenario. We start by accessing the performance of different multimodal generative models in encoding and generating information with missing modalities. In addition, we evaluate the quality of the word samples generated by the robot against human calligraphy in a large-scale user study. The results show that our approach can robustly map multimodal commands to generate accurate handwritten word samples, regardless of the set of modalities used to pass information to the agent.

In summary, the main contributions of this work are:

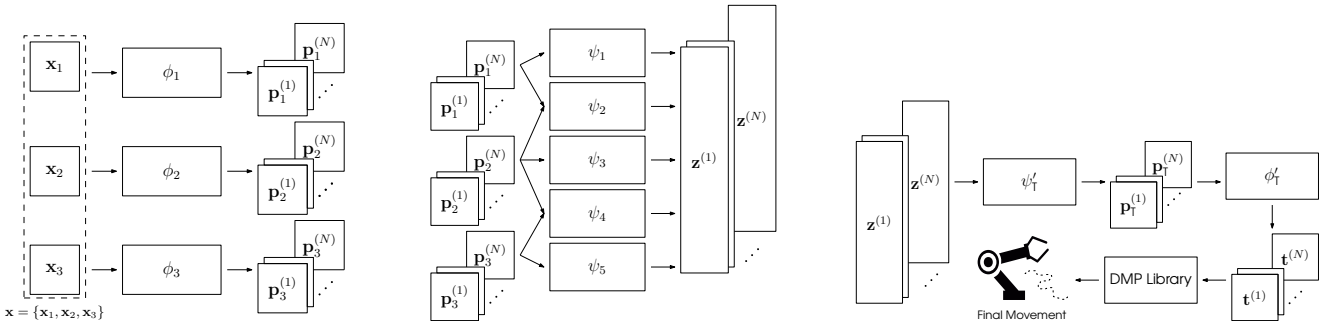
- We propose a novel three-stage pipeline *Perceive-Represent-Generate* (PRG) that translates multimodal information provided by a human user to an adequate movement executed by a robot. Crucially, such mapping is *robust* to missing modality information, as the human may not always provide information through all available communication channels;
- We instantiate our PRG approach in a novel *Robotic Dictaphone* scenario where textual information is converted to robotic motion trajectory, mimicking human handwriting. Our results show that, regardless of the communication channel employed by the human user (e.g., speech or image), our pipeline can accurately translate such information to generate coherent and high-quality handwritten samples.

## II. BACKGROUND

In this work we employ deep generative models to encode information provided by a human user. Of relevance, the variational auto-encoder (VAE) [9] learns to encode latent representations, $\mathbf{z}$, of high-dimensional input data, $\mathbf{x}$, without

*All authors are with INESC-ID & Instituto Superior Técnico, University of Lisbon, Portugal
[1]Corresponding author: fabiovital@tecnico.ulisboa.pt

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981788




(a) Perceive (b) Represent (c) Generate

Fig. 1. The Perceive-Represent-Generate (PRG) framework for multimodal perception and actuation. Details in Section III.

supervision. We assume that the generative process for the input data is $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$ and for the latent space is $\mathbf{z} \sim p(\mathbf{z})$, where the prior $p(\mathbf{z})$ is usually a unit Gaussian distribution. The training of the VAE maximizes the evidence lower bound (ELBO) of the observed data $\mathbf{x}$,

$$\mathscr{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathbb{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})),$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is a neural network parameterized by $\theta$ (*decoder*), and $q_\phi(\mathbf{z}|\mathbf{x})$ is a neural network parameterized by $\phi$ (*encoder*). A recent extension of the VAE framework to the multimodal setting is the Multimodal Unsupervised Sensing (MUSE) model [10]. MUSE employs an hierarchy of representations to learn modality-specific and multimodal latent representations, being robust to missing modality information and scalable to a large number of modalities.

## III. METHODOLOGY

We contribute with *Perceive-Represent-Generate* (PRG), a novel three-stage framework that allows the encoding and generation of high-dimensional multimodal information provided by a human user. As depicted in Fig. 1, in this work, we instantiate PRG in the context of motion trajectory generation for robotic manipulators.

### A. Perceive

We assume that the robot is provided with $M$ sensors to perceive the user command, defining a perceptual space $\mathscr{X} = X_1 \times X_2 \times \ldots \times X_M$. As the user might not employ all available communication channels during task execution, the robot may not be provided with a complete command, $\mathbf{x} \in \mathscr{X}$, but only with a partial view of that command.

To reduce the complexity of the high-dimensional data and remove task-irrelevant information we preprocess the input command $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, discarding unavailable modalities. We define $M$ perceptual maps $\Phi = \{\phi_1, \ldots, \phi_M\}$, responsible for processing and fragmenting the available modality-specific command $\mathbf{x}_m \in X_m$ into a sequence of $N$ individual sub-commands, $\phi_m : \mathbf{x}_m \mapsto \left(\mathbf{p}_m^{(1)}, \ldots, \mathbf{p}_m^{(N)}\right)$. After processing each available modality-specific input, we collect, for each sub-command, the final processed data $\mathbf{p}^{(n)} = \left\{\mathbf{p}_1^{(n)}, \ldots, \mathbf{p}_M^{(n)}\right\}$, where $n \in \{1, \ldots, N\}$.

PRG is agnostic to the nature and number of the perception maps defined by the user for each specific task. Moreover, different maps can be employed to the same modality: raw sound can be encoded into a low-dimensional representation using an encoder or decomposed into label information employing a pre-trained speech-to-text model. In addition, identity perceptual maps can also be easily defined, returning the input as a sequence with one element $\left(\mathbf{p}_m^{(1)}\right) = \mathbf{x}_m$.

### B. Represent

In this stage, we iteratively encode the sub-commands $\mathbf{p}^{(n)}$ into a joint latent space $\mathscr{Z}$ resulting into a sequence of $N$ latent representations $\mathbf{z}^{(n)}$, where $\mathbf{z}^{(n)} \in \mathscr{Z}$.

The encoding process employs a set of $L$ representation maps $\Psi = \{\psi_1, \ldots, \psi_L\}$, with $L \leq 2^M - 1$ to consider all possible combinations of modalities. The map $\psi_l : \left\{\mathbf{p}_{l_1}^{(n)}, \ldots, \mathbf{p}_{l_K}^{(n)}\right\} \mapsto \mathbf{z}^{(n)}$ sequentially maps the sub-commands from the corresponding subset of available modalities into a multimodal latent representation $\mathbf{z}^{(n)}$, where $\{l_1, \ldots, l_K\} \in \mathscr{P}(\{1, \ldots, M\})$, $K \leq M$, and $\mathscr{P}$ is the powerset function. Additionally, we define $M$ generation maps $\Psi' = \{\psi'_1, \ldots, \psi'_M\}$, where each map $\psi'_m : \mathbf{z}^{(n)} \mapsto \mathbf{p}_m^{(n)}$ allows the generation of modality-specific data $\mathbf{p}_m^{(n)}$ from the corresponding joint latent representation $\mathbf{z}^{(n)}$. The representation and generation maps can be instantiated as the encoders and decoders, respectively, of a multimodal VAE (mVAE) model and can be learned by employing a task-specific dataset prior to task execution.

### C. Generate

PRG can generate any input modality from a joint latent representation since we have a generation map (decoder) for each input modality. In this work, we instantiate PRG for the generation of motion trajectories suitable for robotic manipulators. Consequently, PRG iteratively decodes the sequence of joint latent representations $\mathbf{z}^{(n)}$ into a sequence of motion sub-commands $\mathbf{p}_{\text{T}}^{(n)}$ employing the target motion generation map $\psi'_{\text{T}} : \mathbf{z}^{(n)} \mapsto \mathbf{p}_{\text{T}}^{(n)}$, where $\psi'_{\text{T}} \in \Psi'$ and $\mathbf{p}_{\text{T}}^{(n)} \in \mathbf{p}^{(n)}$.

Subsequently, a final processing map $\phi'_{\text{T}} : \mathbf{p}_{\text{T}}^{(n)} \mapsto \mathbf{t}^{(n)}$ is applied that, if required by the task or robotic platform, allows the transformation of the raw generated trajectories

TABLE I

LOG-LIKELIHOOD METRICS FOR DIFFERENT MVAE MODELS USING THE AUGMENTED "UJI CHAR PEN 2" TEST DATASET. HIGHER IS BETTER.

| | $\log p(\mathbf{x}_\mathrm{T})$ | $\log p(\mathbf{x}_\mathrm{S})$ | $\log p(\mathbf{x}_\mathrm{I})$ | $\log p(\mathbf{x}_\mathrm{T}\|\mathbf{x}_\mathrm{S})$ | $\log p(\mathbf{x}_\mathrm{T}\|\mathbf{x}_\mathrm{I})$ | $\log p(\mathbf{x}_\mathrm{S}\|\mathbf{x}_\mathrm{T})$ | $\log p(\mathbf{x}_\mathrm{I}\|\mathbf{x}_\mathrm{T})$ |
|---|---|---|---|---|---|---|---|
| $\mathrm{PRG}_{\mathrm{CVAE}}(\mathbf{x}_\mathrm{T},\mathbf{x}_\mathrm{S})$ | - | - | - | -192.75 | - | - | - |
| $\mathrm{PRG}_{\mathrm{AVAE}}(\mathbf{x}_\mathrm{T},\mathbf{x}_\mathrm{S})$ | -197.55 | -4.17 | - | -189.28 | - | 1.94 | - |
| $\mathrm{PRG}_{\mathrm{AVAE}}(\mathbf{x}_\mathrm{T},\mathbf{x}_\mathrm{I})$ | -197.69 | - | -743.57 | - | -186.28 | - | -730.44 |
| $\mathrm{PRG}_{\mathrm{MUSE}}(\mathbf{x}_\mathrm{T},\mathbf{x}_\mathrm{S},\mathbf{x}_\mathrm{I})$ | -198.04 | -4.53 | -742.49 | -198.10 | -193.63 | 1.96 | -735.02 |

$\mathbf{p}_\mathrm{T}^{(n)}$ into transformed motion trajectories $\mathbf{t}^{(n)}$. We find it advantageous to have this final processing map to create more complex trajectories: $\phi'_\mathrm{T}$ allows the transformation (e.g., sizing, translation) of each individual trajectory before merging them. Finally, all transformed motion sub-commands are concatenated and converted into a single DMP [11] ready to be executed by the robotic agent.

## IV. ROBOTIC DICTAPHONE

We introduce the *Robotic Dictaphone* scenario, where the robot's goal is to generate handwritten word samples from information provided by the human user through three different communication channels $\mathbf{x} = \{\mathbf{x}_\mathrm{T}, \mathbf{x}_\mathrm{S}, \mathbf{x}_\mathrm{I}\}$, where $\mathbf{x}_\mathrm{T}$ and $\mathbf{x}_\mathrm{I}$ are a sequence of 2D letter trajectories and letter images that compose the word, respectively, and $\mathbf{x}_\mathrm{S}$ is a sound corresponding to the word. At execution time, the human user may only employ a subset of such communication channels to provide the words. Therefore, PRG must learn to encode a multimodal representation robust to potential missing modality information. We now describe each of the PRG stages for this scenario.

### A. Perceive

The incoming raw observation data is processed and decomposed into individual sub-commands, in this case, the letters of the word. We define the perception maps specific to each modality, $\Phi = \{\phi_\mathrm{T}, \phi_\mathrm{S}, \phi_\mathrm{I}\}$. For the motion and image modalities, $\phi_\mathrm{T}$ and $\phi_\mathrm{I}$, we return the given sequence after normalizing each of its elements, trajectories and images, respectively. For the sound perception map, $\phi_\mathrm{S}$, we employ wav2vec 2.0, a self-supervised learning framework for speech recognition [12]. Hence, we process the raw audio data into the label information associated with each letter, allowing for a more efficient downstream representation.

### B. Represent

As explained in Section III-B, PRG is agnostic to the mVAE model employed. The *Represent* stage of PRG can be seen as an abstraction of mVAE's encoding phase. Depending on the mVAE model used, the PRG's representation maps will differ to accommodate the encoding channels provided by the model.

In Section V-A we evaluate the performance of PRG instantiated with different mVAE models. We train all mVAE models on data provided from the UJI Char Pen 2 dataset[1], from which we only select one-stroke formed digits and letters. We further augment the dataset by sampling from a probabilistic model derived for each character, following the procedure of [13].

### C. Generate

Similarly to the *Represent* stage, the PRG's *Generate* stage is an abstraction of the decoding phase of the mVAE model. In particular, we consider the motion generation map $\psi'_\mathrm{T}$, provided by the mVAE to generate trajectory samples given the latent representation. Furthermore, we define the final processing map $\phi'_\mathrm{T}$ to homogenize the generated trajectories for each letter, following: 1) We scale all trajectories appropriately to their expected proportion in the final word regarding a predefined heuristic (e.g., lowercase `a` should have half-height of an uppercase `A`); 2) We translate all trajectories vertically, accordingly to the heuristic that every letter must start at the origin, except for {`f`, `g`, `j`, `p`, `q`, `y`} which begin at a lower predefined coordinate; 3) We define a fixed horizontal distance between two consecutive trajectories; 4) Generate *connection* trajectories[2] between the end and the beginning of two consecutive letters.

After applying $\phi'_\mathrm{T}$, we concatenate the $N$ letter trajectories and $N-1$ connection trajectories and convert them into a single DMP for the target word, executable by the robot.

## V. EVALUATION

We evaluate our PRG framework in the *Robotic Dictaphone* scenario. Firstly, we quantitatively evaluate the generative capability of different mVAE models integrated into PRG. Secondly, we evaluate qualitatively the handwritten samples generated by PRG considering different input modalities. Finally, we assess the quality of PRG samples against human handwriting in a large-scale user study.

### A. Quantitive Evaluation of mVAE models for PRG

We employ and compare several mVAE models to learn the representation and generation maps $\Psi$ and $\Psi'$, respectively, required to encode multimodal data and generate the target motion trajectories. To evaluate the robustness of the *Represent* and *Generate* stages, we consider four different mVAE models:

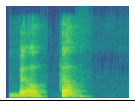
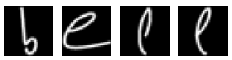
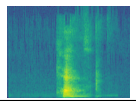
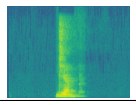
1) $\mathrm{PRG}_{\mathrm{CVAE}}(\mathbf{x}_\mathrm{T}, \mathbf{x}_\mathrm{S})$: We employ the CVAE model [14] to learn the set of maps $\Psi = \{\psi_{\mathrm{T,S}}, \psi_\mathrm{S}\}$ ($\psi_{\mathrm{T,S}}$ is only used at training time) and $\Psi' = \{\psi'_\mathrm{T}\}$ in order to generate motion information conditioned on sound information.

[1]To the best of our knowledge, this is the only dataset with all required modalities for English characters.

[2]For more details regarding the final processing map refer to the extended version of the paper available at https://arxiv.org/abs/2204.03051

TABLE II

TRAJECTORY SAMPLES RETRIEVED FROM RUNNING $\text{PRG}_{\text{MUSE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S}, \mathbf{x}_\text{I})$, IN THE ROBOTIC DICTAPHONE SCENARIO, WHEN GIVEN AS INPUT THE SOUND OF THE RESPECTIVE WORD, $\mathbf{x}_\text{S}$ (SPECTROGRAM OF THE RECORDED SOUND SHOWN), OR THE IMAGE OF EACH LETTER OF THE WORD, $\mathbf{x}_\text{I}$.

| | bell | | cat | | jump | |
|---|---|---|---|---|---|---|
| | Sound ($\mathbf{x}_\text{S}$) | Image ($\mathbf{x}_\text{I}$) | Sound ($\mathbf{x}_\text{S}$) | Image ($\mathbf{x}_\text{I}$) | Sound ($\mathbf{x}_\text{S}$) | Image ($\mathbf{x}_\text{I}$) |
| Input |  |  |  |  |  |  |
| Output |  |  |  |  |  |  |

2) $\text{PRG}_{\text{AVAE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S})$: we employ the AVAE model [15] to learn the set of maps $\Psi = \{\psi_\text{T}, \psi_\text{S}\}$ and $\Psi' = \{\psi'_\text{T}, \psi'_\text{S}\}$ in order to encode and generate motion and sound information.
3) $\text{PRG}_{\text{AVAE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{I})$: we employ the AVAE model to learn the set of maps $\Psi = \{\psi_\text{T}, \psi_\text{I}\}$ and $\Psi' = \{\psi'_\text{T}, \psi'_\text{I}\}$ in order to encode and generate motion and image information.
4) $\text{PRG}_{\text{MUSE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S}, \mathbf{x}_\text{I})$: we employ the MUSE model to learn a joint representation map $\Psi = \{\psi_{\text{T,S,I}}\}$, robust to missing modality information, and $\Psi' = \{\psi'_\text{T}, \psi'_\text{S}, \psi'_\text{I}\}$ to generate modality-specific information.

We evaluate the generative performance of all mVAE solutions quantitatively. In Table I, we present standard log-likelihood metrics regarding the marginal and conditional log-likelihoods that are estimated resorting to 1000 and 5000 importance-weighted samples, respectively. The results show no significant benefit of $\text{PRG}_{\text{CVAE}}$, as it is outperformed by $\text{PRG}_{\text{AVAE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S})$ regarding the conditional log-likelihood $\log p(\mathbf{x}_\text{T}|\mathbf{x}_\text{S})$. As for the $\text{PRG}_{\text{MUSE}}$ and both $\text{PRG}_{\text{AVAE}}$ models, the results highlight a compromise between the generative performance and scalability of the approaches: both instances of $\text{PRG}_{\text{AVAE}}$ outperform $\text{PRG}_{\text{MUSE}}$ in terms of learning a quality trajectory representation, $\log p(\mathbf{x}_\text{T})$, and of conditionally generating trajectory information, $\log p(\mathbf{x}_\text{T}|\mathbf{x}_\text{I})$ and $\log p(\mathbf{x}_\text{T}|\mathbf{x}_\text{S})$. However, this approach is not practically extendable to more than two modalities since it needs a new encoder for each combination of modalities. On the other hand, $\text{PRG}_{\text{MUSE}}$ scales linearly with the number of modalities and can learn a joint representation of all modalities, suitable to generate coherent motion information. In order to be able to consider all perceptual modalities as input, we employ $\text{PRG}_{\text{MUSE}}$ throughout the rest of this work.

### B. Qualitative Evaluation of PRG Samples

We start by qualitatively evaluating the performance of PRG in generating handwritten word samples from image and sound information. In Table II, we observe that $\text{PRG}_{\text{MUSE}}$ allows for the generation of coherent and varied word samples, regardless of the input modality employed.

Additionally, we highlight how the PRG framework can be incorporated into standard robotic platforms by considering a simulated environment where the robotic agent executes the final generated trajectory, provided by PRG. We use OpenRAVE [16] as our simulation environment, where we place a dual 7-DOF Baxter robot in front of a table. PRG generates a single handwriting motion from an input modality of a word ($\mathbf{x}_\text{S}$ or $\mathbf{x}_\text{I}$). The generated motion is further transformed into a joint-space trajectory using Baxter's default inverse kinematics procedure before execution. The simulation depicts how PRG allows a robot platform to convert high-dimensional inputs, such as sound, to effective motion trajectories of words, shown in Fig. 2.

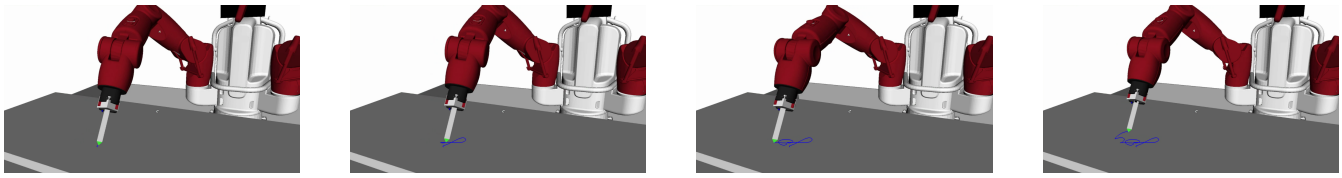
### C. User Study on PRG against Human Calligraphy

We conduct an online user study evaluating the performance of PRG in generating human-like handwritten words. The study implements a Turing-like test approach where the participants have to distinguish the origin of the word sample: human or PRG.

We start with two study hypotheses: (**H1**) the participant cannot distinguish motions handwritten by humans and PRG; (**H2**) the participant will not show high confidence when asked to distinguish handwritten words by humans and PRG. With **H1**, we expect PRG to produce handwritten words similar to human handwriting: we quantify this hypothesis using $\|\hat{c} - c\| \leq \delta$, where $\hat{c}$ and $c$ denote the classification performance from the study and a random guess, respectively, and $\delta$ is a threshold of equivalence. For **H2**, we expect each participant to exhibit low confidence (below the middle confidence value) in their choices, further asserting the subjective similarity of the human and PRG samples.

The study involves two phases: in a first phase, we ask 10 random participants (group 1) to write 10 words in a cursive movement (without lifting the pen). Additionally we employ PRG and generate the same words from label information. In the second phase, another group of 50 participants (group 2) answers an online and anonymous questionnaire using the Prolific platform. For each word, the participants answer two questions: in the first question, we present four randomly written words by group 1 and one by PRG, and we ask the participant to select the word written by PRG. For the second question, we ask the participant to categorize its confidence in the selection (very low, low, neutral, high, and very high).

In order to test **H1**, we subject the corresponding first question of each word to a binomial distribution. Our analysis showed that, on average, the participants achieved a probability of $\hat{c} = 0.422 \pm 0.057$ for choosing the word written by



(a) (b) (c) (d)

Fig. 2. A dual 7-DOFs Baxter manipulator writing the word "joy" in a simulation environment on OpenRAVE. The word motion was dervived by $\text{PRG}_{\text{MUSE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S}, \mathbf{x}_\text{I})$ in the context of the *Robotic Dictaphone* scenario.

(a) (b) (c) (d)

Fig. 3. Example of words handwritten by humans (a), (b), and by $\text{PRG}_{\text{MUSE}}(\mathbf{x}_\text{T}, \mathbf{x}_\text{S}, \mathbf{x}_\text{I})$, in the context of the *Robotic Dictaphone* scenario, (c), (d).

PRG, which is far from a random guess, $c = 0.2$. Two one-sided T-tests further supported this conclusion since it failed to reject one of the null hypotheses, in this case, $\hat{c} < c - \delta$ where $\delta \approx 0.057$ is defined according to the confidence interval of the results obtained, with a probability threshold $\alpha = 0.05$. Such result can be understood due to the differences in appearance between the words generated by PRG and by humans: human samples appear wavier, with more fluctuations throughout the trace than the ones formed by PRG. Furthermore, as shown in Fig. 3, most human samples contain redundant strokes, as we forced their motion to be cursive: for example, the trace for the letter "o" frequently contained two full circles so the participant could adjust the motion making it easier to write the following letter. In contrast, PRG does not display this redundant behaviour since the motion of each letter is learned independently and the final word motion is converted into a single DMP.

Regarding **H2**, the average confidence level was $3.17 \pm 1.5$. One-sided t-test rejected that the average confidence is below neutral, 3, for $\alpha = 0.05$. The participants showed reasonable confidence about their choice, further indicating the differences between words written by humans and PRG.

To further understand the previous results, we ran a second anonymous and online user study, where we showed cursive handwritten words to a new pool of 30 participants, asking them to type the words they observed (the same from the previous study). We showed one group (half the participants) five words written by humans and the other five by PRG. For the second group (other half), we showed the same words from the opposing source, exchanging human with PRG, and vice-versa. We hypothesize that the participants are able to identify the handwritten words regardless if it was written by a human or by PRG. To test this hypothesis, we separate the results from words handwritten by humans from those by PRG. The success rates are $0.78 \pm 0.092$ and $0.83 \pm 0.083$, respectively. We declare that both means are equal for the null hypothesis. Using a two-sided T-test, we reject the alternative hypothesis for the threshold probability $\alpha = 0.05$, meaning that we accept the null hypothesis. Thus, we can state that words generated by PRG are equally readable as human handwriting and the difference observed in the results of the previous user study is due to stylistic traits of the words and not fundamental, semantic, ones.

## VI. RELATED WORK

Several works focus on controlling robots with commands provided by human users through different modalities. In robotic navigation tasks, the use of directional voice commands has been explored to improve the performance of the robot [17]. Another approach considers the uncertainty in the voice commands to facilitate learning [18]. However, most approaches consider a single perceptual modality, often sound, to provide commands. PRG is able to consider multiple modalities to provide commands to a robotic platform.

Other works integrate multiple modalities in order to infer the desired command. A recent work captures audio and visual samples independently, converting them into scores to combine them and determine the command from a known set [19]. Similarly, other approaches employ neural networks to compute the confidence scores for each possible command [20]. These scores can only classify input modalities into predefined label commands. Meanwhile, since PRG learns a joint latent representation of all modalities, it can directly generate any modality as the output command.

Other approaches learn multimodal representations to account for commands provided through multiple input modalities. One method integrates motor and sensory time-series data (motion, image, and sound) in a fused multimodal representation, employing auto-encoder (AE) models [21]. This framework can perform cross-modal retrieval, however, since it employs a standard AE is unable to generate novel instances. Another approach introduced a multimodal architecture for cross-modal inference on visual and bathymetric data [22]. The framework employs a hierarchy of denoising AEs for each modality and a mixture of restricted Boltzman machines (RBM) to learn a multimodal representation. However, training a multimodal representation through RBMs is computationally expensive and prone to divergence, requiring a meticulous model design. Contrary to both approaches, PRG can generate novel instances, through a computationally stable training mechanism (employing an mVAE).

Other approaches employ mVAEs to learn latent representations of multimodal data for robotic tasks. The CVAE model, employed as a baseline in Section V-A, has been successfully integrated in the architecture of robotic agents for pouring and dabbing tasks [14]. The limitation of CVAE is that the generative process is fixed requiring all conditional modalities to generate the target ones. PRG employs MUSE, which, on the other hand, learns a joint latent space where the conditional and target modalities can be any set of modalities. Another approach considers the AVAE model, also employed as a baseline in Section V-A, to write single letters from visual information [15]. While AVAE can also learn a joint latent space, it needs a different encoder for each combination of modalities. In contrast, PRG employs MUSE, which increases linearly with the number of modalities.

## VII. CONCLUSION

In this work, we addressed the problem of *translating* multimodal commands provided by different communication channels to a sequence of movements executed by a robotic agent. We contributed with a novel three-stage pipeline (PRG) that allows the processing, mapping, and generation of trajectory information, regardless of the communication channels employed by the human user. At the core of our pipeline, we leverage multimodal generative models to learn a low-dimensional representation of the high-dimensional data provided by the human, robust to partial observations. We instantiate our pipeline in the context of a novel multimodal robotic handwriting task called the *Robotic Dictaphone* scenario. Our results show that our approach allows the generation of coherent and high-quality handwritten samples, regardless of the human user's communication channels. Furthermore, since PRG can generate any modality from the learned joint latent space, we can easily revert the robotic handwriting task (given the movement to write a word, we can use our framework to output an image or sound corresponding to the given word), as PRG is agnostic to the nature of the downstream task. Additionally, we plan on extending PRG to scenarios with large number of modalities, to test its scalability, and to consider other human-robot interaction tasks, to evaluate the role of multimodal command mapping in such scenarios.

## ACKNOWLEDGEMENTS

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under projects UIDB/50021/2020 (INESC-ID multi-annual funding), PTDC/CCI-COM/7203/2020 (HOTSPOT), and PTDC/CCI-COM/5060/2021 (RELEvaNT). In addition, this work is supported by the EU project TAILOR, under GA No 952215. Miguel Vasco acknowledges the FCT PhD grant SFRH/BD/139362/2018.

## REFERENCES

[1] J. Ruiz-del-Solar, P. Loncomilla, and N. Soto, "A survey on deep learning methods for robot vision," *CoRR*, vol. abs/1803.10862, 2018.

[2] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *Journal of Machine Learning Research*, vol. 22, no. 30, pp. 1–82, 2021.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[4] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[5] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research," *Annual review of biomedical engineering*, vol. 14, pp. 275–294, 2012.

[6] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[7] T. Xue, W. Wang, J. Ma, W. Liu, Z. Pan, and M. Han, "Progress and prospects of multimodal fusion methods in physical human–robot interaction: A review," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 355–10 370, 2020.

[8] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human–robot collaboration," *Autonomous Robots*, vol. 42, no. 5, pp. 957–975, 2018.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] M. Vasco, H. Yin, F. S. Melo, and A. Paiva, "How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '22. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2022, p. 1301–1309.

[11] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 02 2013.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] H. Yin, P. Alves-Oliveira, F. S. Melo, A. Billard, and A. Paiva, "Synthesizing robotic handwriting motion by learning from human demonstrations," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16, 2016, p. 3530–3537.

[14] Y. Hristov and S. Ramamoorthy, "Learning from demonstration with weakly supervised disentanglement," in *International Conference on Learning Representations*, 2021.

[15] H. Yin, F. Melo, A. Billard, and A. Paiva, "Associate latent encodings in learning from demonstrations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 3848–3854, Feb. 2017.

[16] R. Diankov, "Automated construction of robotic manipulation programs," Ph.D. dissertation, Carnegie Mellon University, Robotics Institute, August 2010.

[17] M. V. J. Muthugala, P. A. S. Srimal, and A. B. P. Jayasekara, "Enhancing interpretation of ambiguous voice instructions based on the environment and the user's intention for improved human-friendly robot navigation," *Applied Sciences*, vol. 7, no. 8, p. 821, 2017.

[18] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "Enhancing human-robot interaction by interpreting uncertain information in navigational commands based on experience and environment," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2915–2921.

[19] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2702–2706.

[20] H. Liu, T. Fang, T. Zhou, and L. Wang, "Towards robust human-robot collaborative manufacturing: Multimodal fusion," *IEEE Access*, vol. 6, pp. 74 762–74 771, 2018.

[21] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721–736, 2014.

[22] D. Rao, M. D. Deuge, N. Nourani–Vatani, S. B. Williams, and O. Pizarro, "Multimodal learning and inference from visual and remotely sensed data," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 24–43, 2017.