

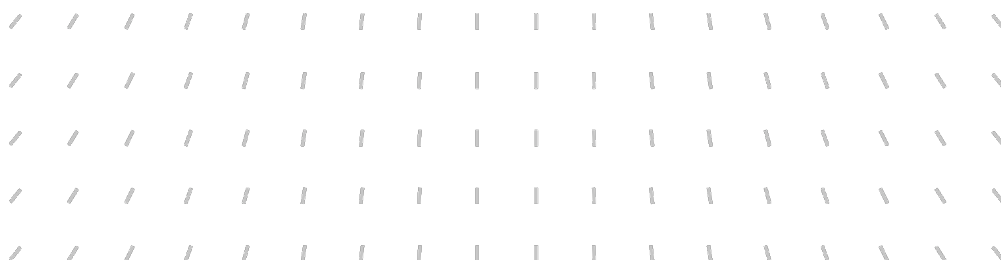
2025 Activity Report

RESEARCH CENTRE: Inria Centre at Université Grenoble Alpes
IN PARTNERSHIP WITH: Université de Grenoble Alpes

Project-Team

ROBOTLEARN

Learning, perception and control for social robots



Project-Team ROBOTLEARN

Creation of the Project-Team: 2021 July 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

Keywords

Computer sciences and digital sciences

- A5.7.3. – Speech
- A5.7.4. – Analysis
- A5.7.5. – Synthesis
- A5.10.2. – Perception
- A5.10.4. – Robot control
- A5.10.5. – Robot interaction (with the environment, humans, other robots)
- A9.2. – Machine learning
- A9.3. – Signal processing
- A9.4. – Natural language processing
- A9.5. – Robotics and AI
- A9.11. – Generative AI
- A9.12.2. – Activity recognition
- A9.12.5. – Object tracking and motion analysis
- A9.14. – Evaluation of AI models

Other research topics and application domains

- B2. – Digital health
- B5.6. – Robotic systems

Contents

Project-Team ROBOTLEARN	1
1 Team members, visitors, external collaborators	5
2 Overall objectives	6
3 Research program	6
3.1 Deep probabilistic models	7
3.2 Human behavior understanding	8
3.3 Learning and control for social robots	9
4 Application domains	11
5 Social and environmental responsibility	13
5.1 Impact of research results	13
6 Highlights of the year	13
6.1 Final results of the H2020 SPRING project	13
6.2 Onboarding of Stéphane Lathuilère	13
6.3 The genesis of ComLearn	14
6.4 Welcome Miroka!	14
7 Latest software developments, platforms, open data	14
7.1 New platforms	14
8 New results	15
8.1 Deep Probabilistic Models	15
8.1.1 Diffusion-based Unsupervised Audio-visual Speech Enhancement	15
8.1.2 No Images, No Problem: Retaining Knowledge in Continual VQA with Questions-Only Memory	15
8.1.3 Group-robust Machine Unlearning	15
8.1.4 DiMO: Distilling Masked Diffusion Models into One-step Generator	16
8.1.5 Don't Forget your Inverse DDIM for Image Editing	16
8.2 Human Behavior Understanding	17
8.2.1 MEGA: Masked Generative Autoencoder for Human Mesh Recovery	17
8.2.2 Unlearning personal data from a single image	17
8.2.3 AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder	17
8.2.4 Posterior Transition Modeling for Unsupervised Diffusion-Based Speech Enhancement	18
8.3 Learning and Control for Social Robots	18
8.3.1 OpenSocInt: A Multi-modal Training Environment for Human-Aware Social Navigation	18
8.3.2 Socially Pertinent Robots in Gerontological Healthcare	18
8.4 Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Two-Wave Comparative Study on Performance, Engagement, and User Perceptions	18
8.5 Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation	19
9 Partnerships and cooperations	19
9.1 International initiatives	19
9.1.1 Inria associate team not involved in an ILL or an international program	19
9.2 International research visitors	20
9.3 European initiatives	21
9.3.1 H2020 projects	21

10 Dissemination	22
10.1 Promoting scientific activities	22
10.1.1 Scientific events: organisation	22
10.1.2 Scientific events: selection	22
10.1.3 Journal	22
10.1.4 Invited talks	22
10.1.5 Leadership within the scientific community	23
10.2 Teaching - Supervision - Juries - Educational and pedagogical outreach	23
10.2.1 Supervision	23
10.2.2 Juries	23
10.2.3 Educational and pedagogical outreach	23
11 Scientific production	23
11.1 Major publications	23
11.2 Publications of the year	25
11.3 Cited publications	27

1 Team members, visitors, external collaborators

Research Scientists

- Xavier Alameda Pineda [Team leader, INRIA, Senior Researcher]
- Laurent Girin [GRENOBLE INP, HDR]
- Patrice Horaud [retired, Emeritus, HDR]
- Thomas Hueber [CNRS]
- Stéphane Lathuilière [INRIA, ISFP]
- Olivier Perrotin [CNRS, Researcher]

Post-Doctoral Fellows

- Xiaoyu Lin [UGA, Post-Doctoral Fellow, until May 2025]
- Samir Sadok [INRIA, Post-Doctoral Fellow, from May 2025]
- Samir Sadok [UGA, Post-Doctoral Fellow, until May 2025]

PhD Students

- Maxime Attwood [UGA, from Oct 2025]
- Gaetan Lepage [INRIA, until Jan 2025]

Technical Staff

- Ahamed Mohamed [INRIA, Engineer]
- Gianluca Zappavigna [INRIA, from Dec 2025]

Interns and Apprentices

- Maxime Attwood [INRIA, Intern, from Feb 2025 until Jul 2025]
- Manal Belouarda [INRIA, Intern, from May 2025 until Jul 2025]
- Gianluca Zappavigna [INRIA, Intern, from Jun 2025 until Nov 2025]

Administrative Assistant

- Nathalie Gillot [INRIA]

Visiting Scientists

- Massimiliano Pappa [UNIV ROME III, until Jul 2025]
- Javier Venema Rodriguez [Panacea Cooperative Research, PhD student at University of Granada, from May 2025 until Jun 2025]

2 Overall objectives

In recent years, social robots have been introduced into public spaces, such as museums, airports, commercial malls, banks, show-rooms, schools, universities, hospitals, and retirement homes, to mention a few examples. In addition to classical robotic skills such as navigating in complex environments, grasping and manipulating objects, i.e. *physical interactions*, social robots must be able to communicate with people and to adopt appropriate behavior. Welcoming newcomers, providing various pieces of information, and entertaining groups of people are typical services that social robots are expected to provide in the near future.

Nevertheless, today's state-of-the-art in robotics is not well-suited to fulfill these needs, and there are two main bottlenecks: (i) robots are limited to a handful of simple scenarios which leads to (ii) social robots not being well accepted by a large percentage of users. While there are research programs and projects which have tackled some of these challenges, existing commercially available robots cannot (or only to a very limited extent) recognize individual behaviors (e.g. facial expressions, hand- and body-gestures, head- and eye-gaze) or group behaviors (e.g. who looks at whom, who speaks to whom, who needs robot assistance, etc.). They do not have the ability to take social (or non-verbal) signals into account while they are engaged in spoken dialogue and they cannot connect the dialogue with the persons and objects that are physically present in their surroundings. We would like to develop robots that are responsible for their perception, and act to enhance the quality of the signals they receive, instead of asking the users to adapt their behavior to the robotic platform.

The scientific ambition of ROBOTLEARN is to train robots to acquire the capacity to **look, listen, learn, move** and **speak** in a socially acceptable manner. We identify three main objectives:

1. Develop deep probabilistic models and methods that allow the fusion of audio and visual data, possibly sequential, recorded with cameras and microphones, and in particular with sensors onboard of robots.
2. Increase the performance of human behaviour understanding using deep probabilistic models and jointly exploiting auditory and visual information.
3. Learn robot-action policies that are socially acceptable and that enable robots to better perceive humans and the physical environment.

ROBOTLEARN stands at the cross-roads of several fields: computer vision, audio signal processing, speech technology, statistical learning, deep learning, and robotics. In partnership with several companies (e.g. PAL Robotics and ERM Automatismes Industriels), the technological objective is to launch a brand new generation of robots that are flexible enough to adapt to the needs of the users, and not the other way around. The experimental objective is to validate the scientific and technological progress in the real world. Furthermore, we believe that ROBOTLEARN will contribute with tools and methods able to process robotic data (perception and action signals) in such a way that connections with more abstract representations (semantics, knowledge) are possible. The developments needed to discover and use such connections could be addressed through collaborations. Similarly, aspects related to robot deployment in the consumer world, such as ethics and acceptability will be addressed in collaboration, for instance, with the Broca day-care hospital in Paris.

From a methodological perspective, the challenge is at least three-fold. First, to reduce the amount of human intervention needed to adapt the designed learning models in a new environment. We aim to further develop strategies based on unsupervised learning and unsupervised domain adaptation, within the framework of deep probabilistic modeling with latent variables [57]. Second, to successfully exploit auditory and visual data for human behavior understanding. For instance by developing mechanisms that manage to model and learn the complementarity between sounds and images [9]. Third, by developing reinforcement learning algorithms that can transfer previous knowledge to future tasks and environments. One potential way forward is to anchor the learning into key features that can be hand-crafted or learned [47].

3 Research program

ROBOTLEARN is structured in three research axes, allowing to develop socially intelligent robots. First, on deep probabilistic models, which include the large family of deep neural network architectures, the large family of probabilistic models, and their intersection. Briefly, we investigate how to jointly exploit the representation power of deep network together with the flexibility of probabilistic models. A well-known

example of such combination are variational autoencoders. Deep probabilistic models are the methodological backbone of the proposed project, and set the foundations of the two other research axes. Second, we develop methods for the automatic understanding of human behavior from both auditory and visual data. To this aim we design our algorithms to exploit the complementary nature of these two modalities, and adapt their inference and on-line update procedures to the computational resources available when operating with robotic platforms. Third, we investigate models and tools allowing a robot to automatically learn the optimal social action policies. In other words, learn to select the best actions according to the social environment. Importantly, these action policies should also allow us to improve the robotic perception, in case this is needed to better understand the ongoing interaction. We believe that these two research axes, grounded on deep and probabilistic models, will ultimately enable us to train robots to acquire social intelligence, meaning, as discussed in the introduction, the capacity to look, listen, learn, move and speak.

3.1 Deep probabilistic models

A large number of perception and interaction processes require temporal modeling. Consider for example the task of extracting a clean speech signal from visual and audio data. Both modalities live in high-dimensional observation spaces and one challenge is to extract low-dimensional embeddings that encode information in a compact way and to update it over time. These high-dimensional to low-dimensional mappings are nonlinear in the general case. Moreover, audio and visual data are corrupted by various perturbations, e.g. by the presence of background noise which is mixed up with the speech signal uttered by a person of interest, or by head movements that overlap with lip movements. Finally, for robotics applications, the available data is scarce, and datasets captured in other settings can only serve as proxies, thus requiring either adaptation [62] or the use of unsupervised models [50]. Therefore, the problem is manifold: to extract low-dimensional compact representations from high-dimensional inputs, to disregard useless data in order to retain information that is relevant for the task at hand, to update and maintain reliable information over time, and to do so in without (or with very few) annotated data from the robot.

This class of problems can be addressed in the framework of state-space models (SSMs). In their most general form, SSMs are stochastic nonlinear systems with latent variables. Such a system is composed of a state equation, that describes the dynamics of the latent (or state) variables, and M observation equations (an observation equation for each sensorial modality m) that predict observations from the state of the system, namely:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{v}_t \quad \mathbf{y}_t^m = g_m(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t^m, \forall m \in \{1 \dots M\}, \quad (1)$$

where the latent vector $\mathbf{x} \in \mathbb{R}^L$ evolves according to a nonlinear stationary Markov dynamic model driven by the observed control variable \mathbf{u} and corrupted by the noise \mathbf{v} . Similarly, the observed vectors $\mathbf{y}^m \in \mathbb{R}^{D_m}$ are modeled with nonlinear stationary functions of the current state and current input, affected by noise \mathbf{w}^m . Models of this kind have been examined for decades and their complexity increases from linear-Gaussian models to nonlinear and non-Gaussian ones. Interestingly, they can also be viewed in the framework of probabilistic graphical models to represent the conditional dependencies between the variables. The objective of an SSM is to infer the sequence of latent variables by computing the posterior distribution of the latent variable, conditioned by the sequence of observations, $p(\mathbf{x}_t | \mathbf{y}_{1:t})$.

When the two functions are linear, the model boils down to a linear dynamical system, that can be learned with an exact Expectation-Maximization (EM) algorithm. Beyond this simple case, non-linearity can be achieved via mixtures of K linear models or more general non-linear (e.g. deep neural) functions. Either case, learning and inference cannot be exact and must be approximated, either by using variational EM algorithms [49, 58, 51, 3], amortized variational inference [57, 45] or a combination of both techniques [46, 18].

We name the larger family of all these methods as Deep Probabilistic Models (DPMs), which form a backbone among the methodological foundations of ROBOTLEARN. Learning DPMs is challenging from the theoretical, methodological and computational points of view. Indeed, the problem of learning, for instance, deep generative Bayesian filters in the framework of nonlinear and non-Gaussian SSMs remains intractable and approximate solutions, that are both optimal from a theoretical point of view and efficient from a computational point of view, remain to be proposed. We plan to investigate both discriminative and generative deep recurrent Bayesian networks and to apply them to audio, visual and audio-visual processing tasks.

Exemplar application: deep probabilistic sequential modeling We have investigated a latent-variable generative model called mixture of dynamical variational autoencoders (MixDVAE) to model the dynamics of a system composed of multiple moving sources. A DVAE model is pre-trained on a single-source dataset to capture the source dynamics. Then, multiple instances of the pre-trained DVAE model are integrated into a multi-source mixture model with a discrete observation-to-source assignment latent variable. The posterior distributions of both the discrete observation-to-source assignment variable and the continuous DVAE variables representing the sources content/position are estimated using the variational expectation-maximization algorithm, leading to multi-source trajectories estimation. We illustrated the versatility of the proposed MixDVAE model on two tasks: a computer vision task, namely multi-object tracking, and an audio processing task, namely single-channel audio source separation. Consequently, this mixture models allows to mix different non-linear source models within the maximum likelihood umbrella and combine the model with other probabilistic models as well.

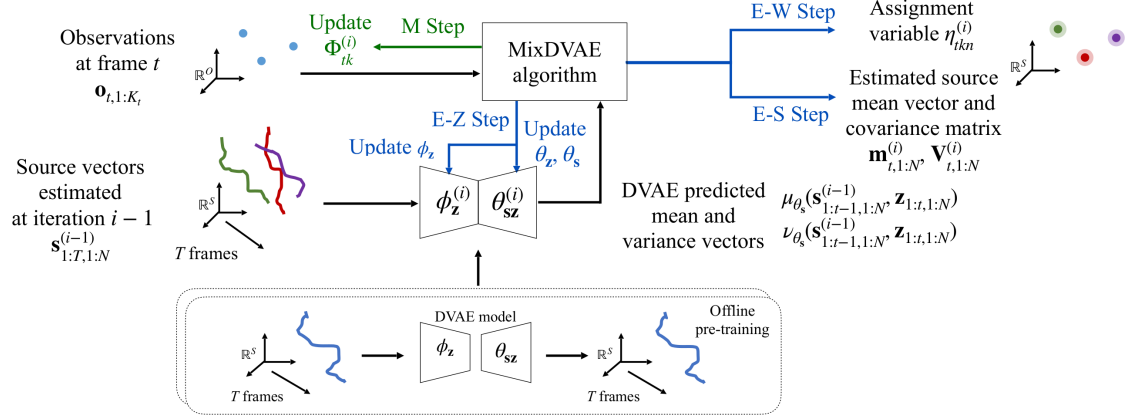


Figure 1: MixDVAE overall diagram.

3.2 Human behavior understanding

Interactions between a robot and a group of people require human behavior understanding (HBU) methods. Consider for example the tasks of detecting eye-gaze and head-gaze and of tracking the gaze directions associated with a group of participants. This means that, in addition to gaze detection and gaze tracking, it is important to detect persons and to track them as well. Additionally, it is important to extract segments of speech, to associate these segments with persons and hence to be able to determine over time who looks to whom and who is the speaker and who are the listeners. The temporal and spatial fusion of visual and audio cues stands at the basis of understanding social roles and of building a multimodal conversational model.

Performing HBU tasks in complex, cluttered and noisy environments is challenging for several reasons: participants come in and out of the camera field of view, their photometric features, e.g. facial texture, clothing, orientation with respect to the camera, etc., vary drastically, even over short periods of time, people look at an object of interest (a person entering the room, a speaking person, a TV/computer screen, a wall painting, etc.) by turning their heads away from the camera, hence facial image analysis is difficult, small head movements are often associated with speech which perturbs both lip reading and head-gaze tracking, etc. Clearly, understanding multi-person human-robot interaction is complex because the person-to-person and person-to-object, in addition to person-to-robot, interactions must explicitly be taken into account.

We propose to perform audio-visual HBU by taking explicitly into account the complementary nature of these two modalities. Differently from one current trend in AV learning [48, 54, 56], we opt for unsupervised probabilistic methods that can (i) assign observations to persons without supervision, (ii) be combined with various probabilistic noise models and (iii) fuse various cues depending on their availability in time (i.e. handle missing data). Indeed, in face-to-face communication, the robot must choose with whom it should engage dialog, e.g. based on proximity, eye gaze, head movements, lip movements, facial expressions, etc., in addition to speech. Unlike in the single-user human-robot interaction case, it is crucial to associate temporal

segments of speech to participants, referred to as speech diarization. Under such scenarios, speech signals are perturbed by noise, reverberation and competing audio sources, hence speech localization and speech enhancement methods must be used in conjunction with speech recognition.

It is also necessary to perform some kind of adaptation to the distribution of the particular data at hand, e.g. collected with robot sensors. If these data are available in advance, off-line adaptation can be done, otherwise the adaptation needs to be performed on-line or at run time. Such strategies will be useful given the particular experimental conditions of practical human-robot interaction scenarios. Either way we will need some sort of on-line learning to perform final adaptation. On-line learning based on deep neural networks is far from being well understood. We plan to thoroughly study the incorporation of on-line learning into both Bayesian and discriminative deep networks. In the practical case of interaction, real-time processing is crucial. Therefore, a compromise must be found between the size of the network, its discriminative power and the computational cost of the learning and prediction algorithms. Clearly, there is no single solution given the large variety of problems and scenarios that are encountered in practice.

Exemplar application: expression-preserving face frontalization Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. We proposed a frontalization methodology that preserves non-rigid facial deformations in order to boost the performance of visually assisted speech communication. The method alternates between the estimation of (i) the rigid transformation (scale, rotation, and translation) and (ii) the non-rigid deformation between an arbitrarily-viewed face and a face model. The method has two important merits: it can deal with non-Gaussian errors in the data and it incorporates a dynamical face deformation model. For that purpose, we used the generalized Student t-distribution in combination with a linear dynamic system in order to account for both rigid head motions and time-varying facial deformations caused by speech production. We proposed to use the zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability of the method to preserve facial expressions. We showed that the method, when incorporated into deep learning pipelines, namely lip reading and speech enhancement, improves word recognition and speech intelligibility scores by a considerable margin.

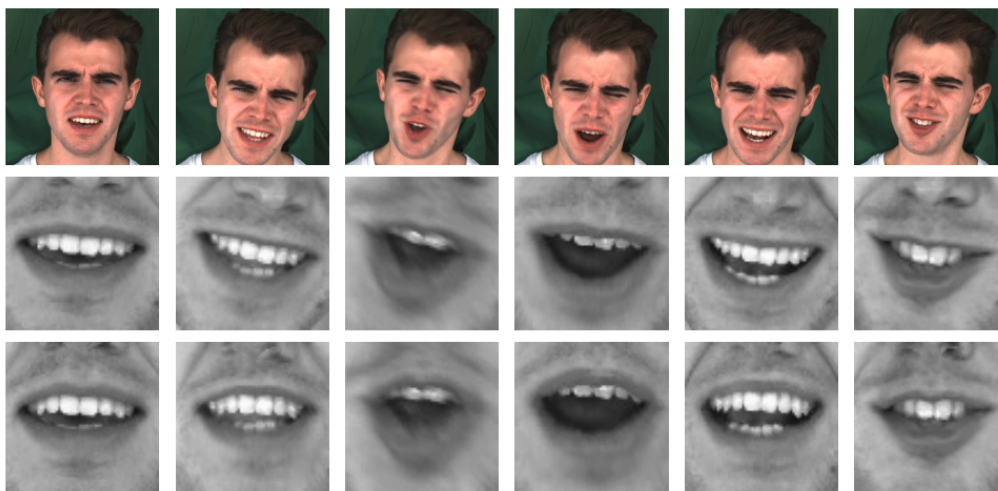


Figure 2: Some results of the proposed expression-preserving face frontalization method.

3.3 Learning and control for social robots

Traditionally, research on human-robot interaction focused on single-person scenarios also called dyadic interactions. However, over the past decade several studies were devoted to various aspects of *multi-party* interactions, meaning situations in which a robot interacts with a group of two or more people [59]. This line of research is much more challenging because of two main reasons. First, the behavioral cues of each individual and of the group need to be faithfully extracted (and assigned to each individual). Second, the behavioral dynamics of groups of people can be pushed by the presence of the robot towards competition [53]

or even bullying [52]. This is why some studies restrict the experimental conditions to very controlled collaborative scenarios, often lead by the robot, such as quiz-like game playing [61] or very specific robot roles [55]. Intuitively, constraining the scenario also reduces the gesture variability and the overall interaction dynamics, leading to methods and algorithms with questionable generalisation to free and natural social multi-party interactions.

Whenever a robot participates in such multi-party interactions, it must perform *social actions*. Such robot social actions are typically associated with the need to perceive a person or a group of persons in an optimal way as well as to take appropriate decisions such as to safely move towards a selected group, to pop into a conversation or to answer a question. Therefore, one can distinguish between two types of robot social actions: (i) *physical actions* which correspond to synthesizing appropriate motions using the robot actuators (motors), possibly within a sensorimotor loop, so as to enhance perception and maintain a natural interaction and (ii) *spoken actions* which correspond to synthesizing appropriate speech utterances by a spoken dialog system. In ROBOTLEARN we will focus on the former, and integrate the latter via collaborations with research groups having with established expertise in speech technologies.

In this regard we face three problems. First, given the complexity of the environment and the inherent limitations of the robot’s perception capabilities, e.g. limited camera field of view, cluttered spaces, complex acoustic conditions, etc., the robot will only have access to a partial representation of the environment, and up to a certain degree of accuracy. Second, for learning purposes, there is no easy way to annotate which are the best actions the robot must choose given a situation: supervised methods are therefore not an option. Third, since the robot cannot learn from scratch by random exploration in a new environment, standard model-free RL approaches cannot be used. Some sort of previous knowledge on the environment or a similar one should be exploited. Finally, given that the robot moves within a populated environment, it is desirable to have the capability to enforce certain constraints, thus limiting the range of possible robot actions.

Building algorithms to endow robots with autonomous decision taking is not straightforward. Two relatively distinct paradigms are available in the literature. First, one can devise customized strategies based on techniques such as *robot motion planning* combined with *sensor-based robot control*. These techniques lack generalization, in particular when the robot acts in complex, dynamic and unconstrained environments. Second, one can let the robot devise its own strategies based on *reinforcement learning* (RL) – a machine learning paradigm in which “agents” learn by themselves by trial and error to achieve successful strategies [60]. It is very difficult, however, to enforce any kind of soft- or hard-constraint within this framework. We will showcase these two scientific streams with one group of techniques for each one: *model predictive control* (MPC) and Q-learning, *deep Q-networks* (DQNs), more precisely. These two techniques are promising. Moreover, they are well documented in the robotics and machine learning. Nevertheless, combining them is extremely challenging.

An additional challenge, independent from the learning and control combination foreseen, is the data distribution gap between the simulations and the real-world. Meta-learning, or the ability to learn how to learn, can provide partial answers to this problem. Indeed, developing machine learning methods able to understand how the learning is achieved can be used to extend this learning to a new task and speed up the learning process on the new task. Recent developments proposed meta-learning strategies specifically conceived for reinforcement learning, leading to Meta-RL methods. One promising trend in Meta-RL is to have a probabilistic formulation involving SSMs and VAEs, i.e. hence sharing the methodology based on dynamical variational autoencoders described before. Very importantly, we are not aware of any studies able to combine Meta-RL with MPC to handle the constraints, and within a unified formulation. From a methodological perspective, this is an important challenge we face in the next few years.

Exemplar application: transferring policies via successor feature representations Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor Representations (SR) and their extension Successor Features (SF) are prominent transfer mechanisms in domains where reward functions change between tasks. They reevaluate the expected return of previously learned policies in a new target task to transfer their knowledge. The SF framework extended SR by linearly decomposing rewards into successor features and a reward weight vector allowing their application in high-dimensional tasks. But this came with the cost of having a linear relationship between reward functions and successor features, limiting its application to tasks where such a linear relationship exists. We proposed a novel formulation of SR based on learning the cumulative discounted probability of successor features, called Successor Feature Representations (SFR). Crucially, SFR allows to reevaluate the

expected return of policies for general reward functions. We introduced different SFR variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on SFR with function approximation demonstrate its advantage over SF not only for general reward functions, but also in the case of linearly decomposable reward functions.

4 Application domains

For the last decades, there has been an increasing interest in robots that cooperate and communicate with people. As already mentioned, we are interested *Socially Assistive Robots* (SARs) that can communicate with people and that are perceived as social entities. So far, the humanoid robots developed to fill this role are mainly used as research platforms for human-robot collaboration and interaction and their prices, if at all commercially available, are in the 6-digit-euro category, e.g. 250,000 EUR for the **iCub robot** and Romeo humanoid robots, developed by the Italian Institute of Technology and SoftBank Robotics Europe, respectively, as well as the **REEM-C** and **TALOS** robots from PAL Robotics. A notable exception being the NAO robot which is a humanoid (legged) robot, available at an affordable price. Apart from humanoid robots, there are also several companion robots manufactured in Europe and available at a much lower price (in the range 10,000–30,000 EUR) that address the SAR market. For example, the **Kompaï**, the **TIAGo**, and the Pepper robots are wheeled indoor robotic platforms. The user interacts with these robots via touch screen and voice commands. The robots manage shopping lists, remember appointments, play music, and respond to simple requests. These affordable robots (Kompaï, TIAGo, NAO, and Pepper) rapidly became the platforms of choice for many researchers in cognitive robotics and in HRI, and they have been used by many EU projects, e.g. **HUMAVIPS**, **EARS**, **VHIA**, and **ENRICHEME**.

When interacting, these robots rely on a few selected modalities. The voice interface of this category of robots, e.g. Kompaï, NAO, and Pepper, is based on speech recognition similar to speech technologies used by smart phones and table-top devices, e.g. Google Home. *Their audio hardware architecture and software packages are designed to handle single-user face-to-face spoken dialogue based on keyword spotting, but they can neither perform multiple sound-source analysis, fuse audio and visual information for more advanced multi-modal/multi-party interactions, nor hold a conversation that exceeds a couple of turns and that is out of very narrow predefined domain.*

To the best of our knowledge, the only notable efforts to overcome some of the limitations mentioned above are the **FP7 EARS** and **H2020 MuMMER** projects. The EARS project's aim was to redesign the microphone-array architecture of the commercially available humanoid robot NAO, and to build a robot head prototype that can support software based on advanced multi-channel audio signal processing. The EARS partners were able to successfully demonstrate the usefulness of this microphone array for speech-signal noise reduction, dereverberation, and multiple-speaker localisation. Moreover, the recent IEEE-AASP Challenge on Acoustic Source Localisation and Tracking (**LOCATA**) comprises a dataset that uses this microphone array. The design of NAO imposed severe constraints on the physical integration of the microphones and associated hardware. Consequently and in spite of the scientific and practical promises of this design, SoftBank Robotics has not integrated this technology into their commercially available robots NAO and Pepper. In order to overcome problems arising from human-robot interaction in unconstrained environments and open-domain dialogue on the Pepper robot, the H2020 MuMMER project aimed to deploy an entertaining and helpful robot assistant to a shopping mall. While they had initial success with short deployments of the robot to the mall, they were not specifically addressing the issues arising from multi-party interaction: Pepper's audio hardware/software design cannot locate and separate several simultaneously emitting speech sources.

To conclude, current robotic platforms available in the consumer market, i.e. with large-scale deployment potential, are neither equipped with the adequate hardware nor endowed with the appropriate software required for multi-party social interactions in real-world environments.

In the light of the above discussion, the partners of the H2020 SPRING project decided to build a robot prototype well suited for socially assistive tasks and shared by the SPRING partners as well as by other EU projects. We participated to the specifications of the ARI robot prototype (shown on the right), designed, developed and manufactured by PAL Robotics, an industrial partner of the SPRING project. ARI is a ROS-enabled, non-holonomic, differential-drive wheeled robot, equipped with a pan and tilt head, with both color and depth cameras and with a microphone array that embeds the latest audio signal processing technologies. Seven ARI robot units were delivered to the SPRING partners in April 2021.



Figure 3: The two robotic platforms of the team: (left) the ARI robot from PAL Robotics and (right) the Miroka robot from EnchantedTools.

We are committed to implement our algorithms and associated software packages onto this advanced robotic platform, from low-level control to high-level perception, interaction and planning tasks, such that the robot has a socially-aware behaviour while it safely navigates in an ever changing environment. We will experiment in environments of increasing complexity, e.g. our robotic lab, the Inria Grenoble cafeteria and Login exhibition, as well as the Broca hospital in Paris. The expertise that the team's engineers and researchers have acquired for the last decade would be crucial for present and future robotic developments and experiments.

5 Social and environmental responsibility

5.1 Impact of research results

Our line of research on developing unsupervised learning methods exploiting audio-visual data to understand social scenes and to learn to interact within is very interesting and challenging, and has large economical and societal impact. Economical impact since the auditory and visual sensors are the most common one, and we can find (many of) them in almost every smartphone in the market. Beyond telephones, manufacturers designing new systems meant for human use, should take into account the need for verbal interaction, and hence for audio-visual perception. A clear example of this potential is the transfer of our technology to a real robotic platform, for evaluation within a day-care hospital (DCH). This is possible thanks to the H2020 SPRING EU project, that assesses the interest of social robotics in the non-medical phases of a regular day for elder patients in a DCH. We are evaluating the performance of our methods for AV speaker tracking, AV speech enhancement, and AV sound source separation, for future technology transfer to the robot manufacturer. This is the first step toward a robot that can be part of the social environment of the DCH, helping to reduce patient and companion stress, at the same time as being a useful tool for the medical personnel. We are confident that developing robust AV perception and action capabilities for robots and autonomous systems, will make them more suitable for environments populated with humans.

6 Highlights of the year

6.1 Final results of the H2020 SPRING project

As the H2020 SPRING project concludes, these joint results highlight the potential of socially assistive robots (SARs) in geriatric care. This research evaluated the humanoid robot ARI in a Paris day-care hospital, focusing on its ability to support older adults and caregivers through multi-modal conversational dialogue. Across several experimental waves involving over 120 participants, the studies assessed system performance, user engagement, and the impact of Large Language Model (LLM) integration. Results from the Acceptability E-Scale (AES) and System Usability Scale (SUS) indicate that end-users are highly receptive to this technology. Key findings demonstrate that while LLMs improve interaction fluency, overall success depends on the robot's ability to minimize errors in cluttered, real-world environments. The study also identified that personal user characteristics and robot adaptability significantly influence long-term adoption and emotional engagement. Ultimately, robust perception and flexible action skills proved essential for moving beyond lab settings into dynamic clinical facilities. These contributions provide a vital framework for deploying AI-driven robotics to alleviate healthcare workloads and reduce patient loneliness. By bridging the gap between technical development and clinical reality, SPRING has paved the way for future geriatric assistive technologies [26, 27].

6.2 Onboarding of Stéphane Lathuilère

A significant milestone in the team's recent evolution was the arrival of Stéphane Lathuilère, who joined as a permanent Research Scientist (ISFP) in January 2025. His integration into RobotLearn—and subsequently ComLearn, see below—brings specialized expertise in deep generative models, image and video generation, and multimodal learning. Having previously served as an Associate Professor at Télécom Paris and completed his PhD within the predecessor Perception team at Inria, Stéphane provides a vital bridge between high-level scene perception and the synthesis of realistic social signals. His research focus on generative AI and "Human

Behavior Understanding" directly supports the new team's mission to develop Multimodal Foundation Models (MFMs). By strengthening the "generation" pillar of the team, his presence accelerates the development of artificial agents capable of more fluid, context-sensitive, and human-centric interactions.

6.3 The genesis of ComLearn

The creation of ComLearn marks a strategic merger between the CRISSP (GIPSA-lab) and RobotLearn (Inria) teams, unifying their world-class expertise in speech synthesis and computer vision. By combining CRISSP's mastery of multimodal generation with RobotLearn's advanced audiovisual perception, ComLearn establishes a powerhouse for next-generation social robotics. This synergy aims to overcome the "last mile" of human-agent interaction by developing Multimodal Foundation Models (MFMs) that ground reasoning and generation in real-world communicative environments. Leveraging a shared methodological foundation in Deep Generative Models, the team will design artificial agents capable of seamless, context-sensitive dialogue within multi-party groups. Beyond technical innovation, the project serves as a bridge between signal processing and cognitive science, providing tools to simulate and better understand fundamental human communication mechanisms. The merger provides the critical mass necessary to lead international research in audio-visual scene analysis and user-adaptive assistive technologies. Ultimately, ComLearn will empower social robots to navigate complex, cluttered social spaces with unprecedented fluency and interpretability.

6.4 Welcome Miroka!

The acquisition of a Miroka robotic platform represents a transformative step for the RobotLearn/ComLearn team, providing a state-of-the-art vehicle for testing Multimodal Foundation Models (MFMs) in real-world settings. Unlike traditional platforms, Miroka's unique globe-based locomotion and "character-driven" design allow it to navigate crowded hospital environments with an agility and social presence that mimics human movement. This platform serves as the ideal physical anchor to ground the team's research in audiovisual perception and generative social signals, bridging the gap between theoretical AI and embodied interaction. Its expressive animated interface provides a high-fidelity canvas for our work in generative behavior synthesis, enabling more nuanced and emotionally resonant communication. Furthermore, Miroka's specialized social capabilities allow the team to study complex multi-party interactions. This investment ensures the team remains at the global forefront of social robotics, moving beyond basic dialogue to truly integrated, context-sensitive assistance. Ultimately, Miroka transforms the lab's algorithmic breakthroughs into tangible, observable social behaviors.

7 Latest software developments, platforms, open data

7.1 New platforms

Participants: Xavier Alameda-Pineda, Ahamed Mohamed, Stéphane Lathuiliere, Nicolas Turro, Soraya Arias.

During 2025, the RobotLearn team has acquired the Miroka platform, see Figure 3 (right). This platform is built by EnchantedTools (a startup in Paris). It has some similarities with our previous platform ARI (that we will keep), namely: the soft appearance, the design intended for social interaction, and multi-sensory capabilities. However, it has some important differences. First, Miroka's face is projected, and therefore much more expressive than the static face of ARI. Second, Miroka comes with integrated LIDAR, which would potentially and significantly help its navigation skills. Third, Miroka moves with a self-balancing strategy over a sphere. While this is more complex to handle, it means that Miroka is a holonomic robot and can move in any direction. We hope it simplifies the issues related to "manouevering" in social settings.

The acquisition of a Miroka robotic platform represents a transformative step for the RobotLearn/ComLearn team, providing a state-of-the-art vehicle for testing Multimodal Foundation Models (MFMs) in real-world settings. Unlike traditional platforms, Miroka's unique globe-based locomotion and "character-driven" design allow it to navigate crowded hospital environments with an agility and social presence that mimics

human movement. This platform serves as the ideal physical anchor to ground the team’s research in audiovisual perception and generative social signals, bridging the gap between theoretical AI and embodied interaction. Its expressive animated interface provides a high-fidelity canvas for our work in generative behavior synthesis, enabling more nuanced and emotionally resonant communication. Furthermore, Miroka’s specialized social capabilities allow the team to study complex multi-party interactions. This investment ensures the team remains at the global forefront of social robotics, moving beyond basic dialogue to truly integrated, context-sensitive assistance. Ultimately, Miroka transforms the lab’s algorithmic breakthroughs into tangible, observable social behaviors.

8 New results

The new results listed below are organised by research axis.

8.1 Deep Probabilistic Models

8.1.1 Diffusion-based Unsupervised Audio-visual Speech Enhancement

Participants: Xavier Alameda-Pineda.

We propose a new unsupervised audiovisual speech enhancement (AVSE) approach that combines a diffusion-based audio-visual speech generative model with a non-negative matrix factorization (NMF) noise model. First, the diffusion model is pre-trained on clean speech conditioned on corresponding video data to simulate the speech generative distribution. This pre-trained model is then paired with the NMF-based noise model to estimate clean speech iteratively. Specifically, a diffusion-based posterior sampling approach is implemented within the reverse diffusion process, where after each iteration, a speech estimate is obtained and used to update the noise parameters. Experimental results confirm that the proposed AVSE approach not only outperforms its audio-only counterpart but also generalizes better than a recent supervised-generative AVSE method. Additionally, the new inference algorithm offers a better balance between inference speed and performance compared to the previous diffusion-based method. Code and demo available [here](#).

8.1.2 No Images, No Problem: Retaining Knowledge in Continual VQA with Questions-Only Memory

Participants: Stéphane Lathuilière.

Continual Learning in Visual Question Answering (VQACL) requires models to learn new visual-linguistic tasks (plasticity) while retaining knowledge from previous tasks (stability). The multimodal nature of VQACL presents unique challenges, requiring models to balance stability across visual and textual domains while maintaining plasticity to adapt to novel objects and reasoning tasks. Existing methods, predominantly designed for unimodal tasks, often struggle to balance these demands effectively. In this work, we introduce QUESION-only replay with Attention Distillation (QUAD), a novel approach for VQACL that leverages only past task questions for regularisation, eliminating the need to store visual data and addressing both memory and privacy concerns. QUAD achieves stability by introducing a question-only replay mechanism that selectively uses questions from previous tasks to prevent overfitting to the current task’s answer space, thereby mitigating the out-of-answer-set problem. Complementing this, we propose attention consistency distillation, which uniquely enforces both intra-modal and inter-modal attention consistency across tasks, preserving essential visual-linguistic associations. Extensive experiments on VQAv2 and NExT-QA demonstrate that QUAD significantly outperforms state-of-the-art methods, achieving robust performance in continual VQA.

8.1.3 Group-robust Machine Unlearning

Participants: Stéphane Lathuilière.

Machine unlearning is an emerging paradigm to remove the influence of specific training data (i.e., the forget set) from a model while preserving its knowledge of the rest of the data (i.e., the retain set). Previous approaches assume the forget data to be uniformly distributed from all training datapoints. However, if the data to unlearn is dominant in one group, we empirically show that performance for this group degrades, leading to fairness issues. This work tackles the overlooked problem of non-uniformly distributed forget sets, which we call group-robust machine unlearning, by presenting a simple, effective strategy that mitigates the performance loss in dominant groups via sample distribution reweighting. Moreover, we present MIU (Mutual Information-aware Machine Unlearning), the first approach for group robustness in approximate machine unlearning. MIU minimizes the mutual information between model features and group information, achieving unlearning while reducing performance degradation in the dominant group of the forget set. Additionally, MIU exploits sample distribution reweighting and mutual information calibration with the original model to preserve group robustness. We conduct experiments on three datasets and show that MIU outperforms standard methods, achieving unlearning without compromising model robustness. Source code available [here](#).

8.1.4 DiMO: Distilling Masked Diffusion Models into One-step Generator

Participants: Stéphane Lathuilière.

Masked Diffusion Models (MDMs) have emerged as a powerful generative modeling technique. Despite their remarkable results, they typically suffer from slow inference with several steps. In this paper, we propose DiMO, a novel approach that distills masked diffusion models into a one-step generator. DiMO addresses two key challenges: (1) the intractability of using intermediate-step information for one-step generation, which we solve through token-level distribution matching that optimizes model output logits by an ‘on-policy framework’ with the help of an auxiliary model; and (2) the lack of entropy in the initial distribution, which we address through a token initialization strategy that injects randomness while maintaining similarity to teacher training distribution. We show DiMO’s effectiveness on both class-conditional and text-conditional image generation, impressively achieving performance competitive to multi-step teacher outputs while drastically reducing inference time. To our knowledge, we are the first to successfully achieve one-step distillation of masked diffusion models and the first to apply discrete distillation to text-to-image generation, opening new paths for efficient generative modeling.

8.1.5 Don’t Forget your Inverse DDIM for Image Editing

Participants: Stéphane Lathuilière.

The field of text-to-image generation has undergone significant advancements with the introduction of diffusion models. Nevertheless, the challenge of editing real images persists, as most methods are either computationally intensive or produce poor reconstructions. This paper introduces SAGE (Self-Attention Guidance for image Editing) - a novel technique leveraging pre-trained diffusion models for image editing. SAGE builds upon the DDIM algorithm and incorporates a novel guidance mechanism utilizing the self-attention layers of the diffusion U-Net. This mechanism computes a reconstruction objective based on attention maps generated during the inverse DDIM process, enabling efficient reconstruction of unedited regions without the need to precisely reconstruct the entire input image. Thus, SAGE directly addresses the key challenges in image editing. The superiority of SAGE over other methods is demonstrated through quantitative and qualitative evaluations and confirmed by a statistically validated comprehensive user study, in which all 47 surveyed users preferred SAGE over competing methods. Additionally, SAGE ranks as the

top-performing method in seven out of 10 quantitative analyses and secures second and third places in the remaining three.

8.2 Human Behavior Understanding

8.2.1 MEGA: Masked Generative Autoencoder for Human Mesh Recovery

Participants: Xavier Alameda Pineda.

Human Mesh Recovery (HMR) from a single RGB image is a highly ambiguous problem, as similar 2D projections can correspond to multiple 3D interpretations. Nevertheless, most HMR methods overlook this ambiguity and make a single prediction without accounting for the associated uncertainty. A few approaches generate a distribution of human meshes, enabling the sampling of multiple predictions; however, none of them is competitive with the latest single-output model when making a single prediction. This work proposes a new approach based on masked generative modeling. By tokenizing the human pose and shape, we formulate the HMR task as generating a sequence of discrete tokens conditioned on an input image. We introduce MEGA, a MaskEd Generative Autoencoder trained to recover human meshes from images and partial human mesh token sequences. Given an image, our flexible generation scheme allows us to predict a single human mesh in deterministic mode or to generate multiple human meshes in stochastic mode. MEGA enables us to propose multiple outputs and to evaluate the uncertainty of the predictions. Experiments on in-the-wild benchmarks show that MEGA achieves state-of-the-art performance in deterministic and stochastic modes, outperforming single-output and multi-output approaches.

8.2.2 Unlearning personal data from a single image

Participants: Stéphane Lathuilière.

Machine unlearning aims to erase data from a model as if the latter never saw them during training. While existing approaches unlearn information from complete or partial access to the training data, this access can be limited over time due to privacy regulations. Currently, no setting or benchmark exists to probe the effectiveness of unlearning methods in such scenarios. To fill this gap, we propose a novel task we call One-Shot Unlearning of Personal Identities (1-SHUI) that evaluates unlearning models when the training data is not available. We focus on unlearning identity data, which is specifically relevant due to current regulations requiring personal data deletion after training. To cope with data absence, we expect users to provide a portraiting picture to aid unlearning. We design requests on CelebA, CelebA-HQ, and MUFAC with different unlearning set sizes to evaluate applicable methods in 1-SHUI. Moreover, we propose MetaUnlearn, an effective method that meta-learns to forget identities from a single image. Our findings indicate that existing approaches struggle when data availability is limited, especially when there is a dissimilarity between the provided samples and the training data. The source code is available [here](#).

8.2.3 AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder

Participants: Samir Sadok, Xavier Alameda Pineda.

This work introduces AnCoGen, a novel method that leverages a masked autoencoder to unify the analysis, control, and generation of speech signals within a single model. AnCoGen can analyze speech by estimating key attributes, such as speaker identity, pitch, content, loudness, signal-to-noise ratio, and clarity index. In addition, it can generate speech from these attributes and allow precise control of the synthesized speech by modifying them. Extensive experiments demonstrated the effectiveness of AnCoGen

across speech analysis-resynthesis, pitch estimation, pitch modification, and speech enhancement. Code and audio examples are available online.

8.2.4 Posterior Transition Modeling for Unsupervised Diffusion-Based Speech Enhancement

Participants: Xavier Alameda Pineda.

We explore unsupervised speech enhancement using diffusion models as expressive generative priors for clean speech. Existing approaches guide the reverse diffusion process using noisy speech through an approximate, noise-perturbed likelihood score, combined with the unconditional score via a trade-off hyperparameter. In this work, we propose two alternative algorithms that directly model the conditional reverse transition distribution of diffusion states. The first method integrates the diffusion prior with the observation model in a principled way, removing the need for hyperparameter tuning. The second defines a diffusion process over the noisy speech itself, yielding a fully tractable and exact likelihood score. Experiments on the WSJ0-QUT and VoiceBank-DEMAND datasets demonstrate improved enhancement metrics and greater robustness to domain shifts compared to both supervised and unsupervised baselines.

8.3 Learning and Control for Social Robots

8.3.1 OpenSocInt: A Multi-modal Training Environment for Human-Aware Social Navigation

Participants: Xavier Alameda-Pineda.

We introduce OpenSocInt, an open-source software package providing a simulator for multi-modal social interactions and a modular architecture to train social agents. We described the software package and showcased its interest via an experimental protocol based on the task of social navigation. Our framework allows for exploring the use of different perceptual features, their encoding and fusion, as well as the use of different agents. The software is already publicly available under GPL [here](#).

8.3.2 Socially Pertinent Robots in Gerontological Healthcare

Participants: Soraya Arias, Nicolas Turro, Alex Auternaud, Chris Reinke, Victor Sanchez, Xavier Alameda-Pineda.

Despite the many recent achievements in developing and deploying social robotics, there are still many underexplored environments and applications for which systematic evaluation of such systems by end-users is necessary. While several robotic platforms have been used in gerontological healthcare, the question of whether or not a social interactive robot with multi-modal conversational capabilities will be useful and accepted in real-life facilities is yet to be answered. This paper is an attempt to partially answer this question, via two waves of experiments with patients and companions in a day-care gerontological facility in Paris with a full-sized humanoid robot endowed with social and conversational interaction capabilities. The software architecture, developed during the H2020 SPRING project, together with the experimental protocol, allowed us to evaluate the acceptability (AES) and usability (SUS) with more than 60 end-users. Overall, the users are receptive to this technology, especially when the robot perception and action skills are robust to environmental clutter and flexible to handle a plethora of different interactions.

8.4 Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Two-Wave Comparative Study on Performance, Engagement, and User Perceptions

Participants: Xavier Alameda Pineda.

Healthcare systems struggle to meet the complex needs of older adults in resource-limited settings. Socially assistive robots (SARs) offer a potential solution by providing information and practical support. This study evaluated the integration of Large Language Models (LLMs) into SARs to improve interaction fluency. Researchers compared a basic dialogue system (Wave 1) to an LLM-based system (Wave 2). The evaluation focused on system performance, interaction success, and multidimensional user engagement. Conducted over eight months in a Paris geriatric hospital, the study involved 28 participants aged 60+. Interactions were video-recorded to code for technical errors and verbal, physical, and emotional engagement. Validated scales were used to measure the robot's overall usability and user acceptability. Results analyzed how user characteristics influenced perceptions of the LLM-enhanced technology. The findings aim to minimize conversational errors and optimize SAR adaptability for real-world use. This research provides insights into successfully deploying AI-driven robotics in geriatric care.

8.5 Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation

Participants: Xavier Alameda Pineda.

Socially assistive robots (SARs) aim to support older adults and clinicians by promoting well-being and managing routine tasks. However, ensuring high levels of acceptability and usability remains a significant hurdle in dynamic care settings. This study evaluated these factors by deploying the ARI humanoid robot in a geriatric day care hospital. Over one year, 97 participants—comprising 65 older patients and 32 informal caregivers—engaged with the robot. The evaluation took place in a waiting area in Paris, where ARI utilized voice-based dialogue for interaction. Researchers employed a mixed-methods approach to capture a holistic view of the user experience. Quantitative data were gathered through the Acceptability E-scale and the System Usability Scale. These assessments were administered orally to accommodate the participants' accessibility needs. Qualitative feedback was also collected to identify subjective perceptions and specific contextual barriers. The study sought to pinpoint key factors influencing the adoption of SARs by both patients and caregivers. Ultimately, the findings provide a framework for improving robot integration into real-world geriatric environments.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria associate team not involved in an IIL or an international program

VisaSpeech

Participants: Xavier Alameda Pineda, Samir Sadok, Stéphane Lathuilière.

Title: Visually Assisted Speech Processing

Duration: 2025 ->

Coordinator: Mirco Ravanelli (mirco.ravanelli@concordia.ca)

Partners:

- Concordia University Montréal (Canada)

Inria contact: Xavier Alameda Pineda

Summary: Fostered by deep learning models trained on massive datasets, artificial intelligence (AI) has recently changed the face of many subfields of computer science and information processing, including speech and audio, computer vision, natural language processing, and robotics. Large language models (LLMs) have become central in modern AI to process the sensory information of the world around us. Originally developed for text, LLMs have now been successfully extended to multimodal signals. Recently, some models for audio-visual speech have also been proposed to learn a joint representation of the clean speech audio signal and the lips images. These models have proven to be very useful for tasks such as audio-visual speech enhancement and recognition. While this research provides valuable insights into exploiting lip-related visual content for speech processing, little is known about foundation models exploiting other visual cues for speech processing. For instance: the speaker's background provides information on the type of environment (e.g. living room, backyard, kitchen), and therefore on the characteristics of the noise, to better guide the enhancement algorithm; the understanding of the surrounding objects could guide the speech recognition model to better infer a missing word; the head orientation could bring insights on how is the current speaker in a conversation. To our knowledge, there is no methodology so far exploiting and/or developing foundation models exploiting lip-unrelated visual cues for speech processing. VisaSpeech will develop models and algorithms to jointly exploit this rich amount of information, thanks to the complementary expertise of the RobotLearn Inria team and Mirco Ravanelli's lab at Concordia University.

9.2 International research visitors

Other international visits to the team

Massimiliano Pappa

Participants: Xavier Alameda Pineda, Stéphane Lathuilière.

Status: PhD

Institution of origin: Università della Sapienza, Roma

Country: Italy

Dates:

Context of the visit: Mobility during PhD

Mobility program/type of mobility: Research Stay

Summary: Deploying safety-critical agents requires anticipating the consequences of actions before they are executed. While world models offer a paradigm for this proactive foresight, current approaches relying on visual simulation incur prohibitive latencies, often exceeding several seconds per step. In this work, we challenge the assumption that visual processing is necessary for safety. We introduce the Latent Sufficiency Hypothesis, positing that a good policy's internal representation, combined with its predicted actions, constitutes a sufficient statistic for predicting the near future observations. To harness this, we present DILLO (Distilled Language-Action World Model), a fast safety layer that shifts the paradigm from "simulate-then-act" to "describe-then-act". Crucially, DILLO creates a "Zero-Visual-Overhead" inference path, bypassing heavy visual encoders entirely. Experiments on MetaWorld tasks demonstrate that DILLO serves as an effective rejection sampling controller.

Javier Venema Rodriguez**Participants:** Stéphane Lathuilière.**Status** PhD**Institution of origin:** Panacea Cooperative Research, Universidad de Granada**Country:** Spain**Dates:** May-June/2025**Context of the visit:****Mobility program/type of mobility:** Research Stay

Summary: Craniofacial reconstruction (CFR) is an identification technique that allows reconstructing facial appearance only from the skull structure, of special relevance in situations where there are no reference data or samples (e.g., medical records, family DNA). The main objective of this work is to develop a reliable and objective method, comparing different strategies based on the use of generative AI, that allow the automation of CFR and its forensic use. With that aim, three strategies have been followed: (i) the use of generative adversarial neural networks (GANs) with volumetric images (3D), (ii) the use of GANs with multi-view depth maps (2.5D) building up on the work of Pan et al. 2024 [1], and (iii) the use of diffusion models. The training has been carried out on a sample with more than a thousand examples sourced from public repositories (NMDID) and collaborations (NFS Seoul), facilitated by access to the computational resources of EuroHPC (MNS 5, BSC).

Preliminary results point to superior performance of 2.5D GANs compared to the rest in terms of quality and fidelity to the real image. Within this approach, the best results so far have been obtained by using three views of the skull model (-30, 0, and 30 degrees) as input, in combination with the use of Wasserstein GAN with gradient penalty (WGAN-GP) in training. In a cross-comparison of CFR outputs and ground truth images, a ranking of correspondence was calculated combining different metrics (MAE and perceptual loss) placing the correct identity in position 4.88 as average. In summary, the use of GANs on 2.5D images constitutes a promising strategy for the development of an automatic CFR tool for forensic use, given that it also offers lower computational costs and environmental impact than other computationally intensive approaches. These results form the basis for future developments towards a photorealistic CFR tool.

9.3 European initiatives

9.3.1 H2020 projects

Participants: Stéphane Lathuilière.**Title:** FaceGEN**Duration:** 1 year (2025)**Coordinator:** Victoria Ulloa (victoria.ulloa@panacea-coop.com)**Partners:**

- Panacea Cooperative Research, Spain
- University of Granada, Spain

Inria contact: Stéphane Lathuilière

Summary: Forensic human identification is an essential step in both criminal investigations and humanitarian efforts. Traditional methods such as DNA profiling, fingerprints, and dental charts are often highly reliable. Still, they depend on the availability of ante-mortem data and the physical condition of the remains. Unfortunately, in many cases, particularly after natural disasters, armed conflicts, or when dealing with remains that are decades old, these methods cannot be applied. In such scenarios, forensic anthropology provides alternative routes. One of these is Craniofacial Reconstruction (CFR), the process of recreating a person’s facial appearance starting from their skull. CFR is based on the well-established correlation between bone structure and soft tissue morphology. Today, however, it remains largely a manual process, requiring the expertise of highly specialized forensic artists. These reconstructions are costly, time-intensive, and difficult to scale. This is where AI and, in particular, generative AI enter the picture. Recent advances in image generation models and high-performance computing resources have opened the door to automating CFR in a way that was unthinkable just a few years ago. By training AI systems to learn from large collections of images, it is now possible to model the relationship between skull shapes and facial features. For forensic practitioners, this will mean faster, more reproducible, and objective reconstructions. For society, it offers new ways to provide closure in unsolved cases and to address the growing number of unidentified remains worldwide.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair As General co-Chair of ACM Multimedia 2026, Xavier Alameda Pineda started working on the organisation of that conference’s edition.

Member of the organizing committees As a web-Chair of ACM Multimedia 2026, Stéphane Lathuilière started working on the website of the conference.

10.1.2 Scientific events: selection

Member of the conference program committees : Xavier Alameda Pineda was Senior Area Chair of ACM Multimedia 2025, and Area Chair of IEEE ICASSP’25 and ICIAP’25.

Stéphane Lathuilière was Area Chair of ICCV 2025 and CVPR 2025

Reviewer : Stéphane Lathuilière was reviewer for WACV 2025 (rounds 1 and 2)

10.1.3 Journal

Member of the editorial boards : during 2025 Xavier Alameda Pineda was Associate Editor of ACM TOMM and CVIU.

Reviewer : Stéphane Lathuilière was reviewer for TMLR

10.1.4 Invited talks

Xavier Alameda Pineda was invited to give a course on the topic “From VAE to Diffusion: probabilistic learning with audio-visual data” at the INPT AI Summer School and an invited talk on “Multimodal perception, action, and evaluation of socially intelligent robots” at the International Workshop on AI for Robotics, organised by Naver Labs Europe.

10.1.5 Leadership within the scientific community

Xavier Alameda Pineda is deeply involved in the multimedia community at the European and International level. At the European level, Xavier is one of the founders of the SIGMM European Chapter, first as Chair (2024-2025), then as Treasurer (2025-2028). At the international level, Xavier is part of the Steering Committee of ACM Multimedia since 2022.

10.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

10.2.1 Supervision

Xavier Alameda Pineda supervised the following PhD students: Gaétan Lepage (defended), Jean-Eudes Ayilo, Sofiene Kammoun, and Maxime Attwood.

Stéphane Lathuilière supervised the following PhD students: Maxime Attwood, Hugo Malard, Sarra Khairi, Imad Marouf, Yasser Benigmim (defended), Thomas De Min, Yuanzhi Zhu.

10.2.2 Juries

Xavier Alameda Pineda was the Chair of the HDR committee of Sergi Pujades, the Chair of the PhD Committee of Timothée Darcet, and of Rim Rekik.

Xavier Alameda Pineda participated in the Selection Committee of the Public Exam for Research Positions at Inria U. Côte d'Azur and of a Maître de Conférences at Télécom ParisTech.

Stéphane Lathuilière was reviewer for the PhD of Paul Couairon and Nicola Dall'Asen.

10.2.3 Educational and pedagogical outreach

Xavier Alameda Pineda participated in two Masters courses: Generative Multimodal AI, and Learning, Probabilities, and Causality. Stéphane Lathuilière participated in a UGA Masters course: "Generative Multimodal AI" and 1 Ensimag course "Perception, Vision et Apprentissage "

11 Scientific production

11.1 Major publications

- [1] L. Airale, D. Vaufray and X. Alameda-Pineda. 'SocialInteractionGAN: Multi-person Interaction Sequence Generation'. In: *IEEE Transactions on Affective Computing* (11th May 2022). DOI: [10.1109/TAFFC.2022.3171719](https://doi.org/10.1109/TAFFC.2022.3171719). URL: <https://hal.inria.fr/hal-03163467>.
- [2] Y. Ban, X. Alameda-Pineda, C. Evers and R. Horaud. 'Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM'. In: *IEEE Signal Processing Letters* 26.6 (1st June 2019), pp. 798–802. DOI: [10.1109/LSP.2019.2908376](https://doi.org/10.1109/LSP.2019.2908376). URL: <https://hal.inria.fr/hal-01969050>.
- [3] Y. Ban, X. Alameda-Pineda, L. Girin and R. Horaud. 'Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.5 (1st May 2021), pp. 1761–1776. DOI: [10.1109/TPAMI.2019.2953020](https://doi.org/10.1109/TPAMI.2019.2953020). URL: <https://hal.inria.fr/hal-01950866> (cit. on p. 7).
- [4] X. Bie, S. Leglaive, X. Alameda-Pineda and L. Girin. 'Unsupervised Speech Enhancement using Dynamical Variational Autoencoders'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30 (16th Sept. 2022), pp. 2993–3007. DOI: [10.1109/TASLP.2022.3207349](https://doi.org/10.1109/TASLP.2022.3207349). URL: <https://hal.inria.fr/hal-03295630>.
- [5] G. Delorme, Y. Xu, S. Lathuilière, R. Horaud and X. Alameda-Pineda. 'CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-Identification'. In: *ICPR 2020 - 25th International Conference on Pattern Recognition*. Milano, Italy: IEEE, 2021, pp. 4428–4435. DOI: [10.1109/ICPR48806.2021.9412431](https://doi.org/10.1109/ICPR48806.2021.9412431). URL: <https://hal.inria.fr/hal-02882285>.

- [6] G. Evangelidis and R. Horaud. ‘Joint Alignment of Multiple Point Sets with Batch and Incremental Expectation-Maximization’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (1st June 2018), pp. 1397–1410. DOI: [10.1109/TPAMI.2017.2717829](https://doi.org/10.1109/TPAMI.2017.2717829). URL: <https://hal.inria.fr/hal-01413414>.
- [7] I. Gebru, S. Ba, X. Li and R. Horaud. ‘Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5 (2nd July 2018), pp. 1086–1099. DOI: [10.1109/TPAMI.2017.2648793](https://doi.org/10.1109/TPAMI.2017.2648793). URL: <https://hal.inria.fr/hal-01413403>.
- [8] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. ‘Dynamical Variational Autoencoders: A Comprehensive Review’. In: *Foundations and Trends in Machine Learning* 15.1-2 (2nd Dec. 2021), pp. 1–175. DOI: [10.1561/22000000089](https://doi.org/10.1561/22000000089). URL: <https://hal.inria.fr/hal-02926215>.
- [9] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. ‘Expression-preserving face frontalization improves visually assisted speech processing’. In: *International Journal of Computer Vision* (12th Jan. 2023). DOI: [10.1007/s11263-022-01742-1](https://doi.org/10.1007/s11263-022-01742-1). URL: <https://hal.science/hal-03902610> (cit. on p. 6).
- [10] S. Lathuilière, B. Massé, P. Mesejo and R. Horaud. ‘Neural Network Based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction’. In: *Pattern Recognition Letters* 118 (1st Feb. 2019), pp. 61–71. DOI: [10.1016/j.patrec.2018.05.023](https://doi.org/10.1016/j.patrec.2018.05.023). URL: <https://hal.inria.fr/hal-01643775>.
- [11] S. Lathuilière, P. Mesejo, X. Alameda-Pineda and R. Horaud. ‘A Comprehensive Analysis of Deep Regression’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (1st Sept. 2020), pp. 2065–2081. DOI: [10.1109/TPAMI.2019.2910523](https://doi.org/10.1109/TPAMI.2019.2910523). URL: <https://hal.inria.fr/hal-01754839>.
- [12] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. ‘Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments’. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (8th Mar. 2019), pp. 88–103. DOI: [10.1109/JSTSP.2019.2903472](https://doi.org/10.1109/JSTSP.2019.2903472). URL: <https://hal.inria.fr/hal-01851985>.
- [13] X. Li, S. Gannot, L. Girin and R. Horaud. ‘Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutional Transfer Function’. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.10 (21st May 2018), pp. 1755–1768. DOI: [10.1109/TASLP.2018.2839362](https://doi.org/10.1109/TASLP.2018.2839362). URL: <https://hal.inria.fr/hal-01645749>.
- [14] X. Li, L. Girin, S. Gannot and R. Horaud. ‘Multichannel Speech Separation and Enhancement Using the Convolutional Transfer Function’. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 27.3 (1st Mar. 2019), pp. 645–659. DOI: [10.1109/TASLP.2019.2892412](https://doi.org/10.1109/TASLP.2019.2892412). URL: <https://hal.inria.fr/hal-01799809>.
- [15] X. Li, S. Leglaive, L. Girin and R. Horaud. ‘Audio-noise Power Spectral Density Estimation Using Long Short-term Memory’. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 918–922. DOI: [10.1109/LSP.2019.2911879](https://doi.org/10.1109/LSP.2019.2911879). URL: <https://hal.inria.fr/hal-02100059>.
- [16] X. Lin, L. Girin and X. Alameda-Pineda. ‘Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation’. In: *Transactions on Machine Learning Research Journal* (2024), pp. 1–19. URL: <https://inria.hal.science/hal-03584014>.
- [17] B. Massé, S. Ba and R. Horaud. ‘Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.11 (1st Nov. 2018), pp. 2711–2724. DOI: [10.1109/TPAMI.2017.2782819](https://doi.org/10.1109/TPAMI.2017.2782819). URL: <https://hal.inria.fr/hal-01511414>.
- [18] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin and R. Horaud. ‘Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders’. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (30th May 2020), pp. 1788–1800. DOI: [10.1109/TASLP.2020.3000593](https://doi.org/10.1109/TASLP.2020.3000593). URL: <https://hal.inria.fr/hal-02364900> (cit. on p. 7).

- [19] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda and R. Ségurier. ‘A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning’. In: *Neural Networks* 172 (Apr. 2024), p. 106120. DOI: [10.1016/j.neunet.2024.106120](https://doi.org/10.1016/j.neunet.2024.106120). URL: <https://inria.hal.science/hal-04132316>.
- [20] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci and N. Sebe. ‘Increasing Image Memorability with Neural Style Transfer’. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 15.2 (1st June 2019). DOI: [10.1145/3311781](https://doi.org/10.1145/3311781). URL: <https://hal.inria.fr/hal-01858389>.
- [21] L. Vaquero, Y. Xu, X. Alameda-Pineda, V. M. Brea and M. Mucientes. ‘Lost and Found: Overcoming Detector Failures in Online Multi-Object Tracking’. In: *ECCV 24 - 18th European Conference on Computer Vision*. Milan (Italie), Italy, 14th July 2024, pp. 1–30. URL: <https://inria.hal.science/hal-04650044>.
- [22] D. Xu, X. Alameda-Pineda, W. Ouyang, E. Ricci, X. Wang and N. Sebe. ‘Probabilistic Graph Attention Network with Conditional Kernels for Pixel-Wise Prediction’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.5 (1st May 2022), pp. 2673–2688. DOI: [10.1109/TPAMI.2020.3043781](https://doi.org/10.1109/TPAMI.2020.3043781). URL: <https://hal.inria.fr/hal-03328687>.
- [23] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus and X. Alameda-Pineda. ‘TransCenter: Transformers With Dense Representations for Multiple-Object Tracking’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (28th Nov. 2022), pp. 1–16. DOI: [10.1109/TPAMI.2022.3225078](https://doi.org/10.1109/TPAMI.2022.3225078). URL: <https://hal.inria.fr/hal-03906940>.
- [24] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, M. Nabi, X. Alameda-Pineda and E. Ricci. ‘Uncertainty-aware Contrastive Distillation for Incremental Semantic Segmentation’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (31st Mar. 2022), pp. 1–14. DOI: [10.1109/TPAMI.2022.3163806](https://doi.org/10.1109/TPAMI.2022.3163806). URL: <https://hal.inria.fr/hal-03908664>.
- [25] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, H. Tang, X. Alameda-Pineda and E. Ricci. ‘Continual Attentive Fusion for Incremental Learning in Semantic Segmentation’. In: *IEEE Transactions on Multimedia* (14th Apr. 2022). DOI: [10.1109/TMM.2022.3167555](https://doi.org/10.1109/TMM.2022.3167555). URL: <https://hal.inria.fr/hal-03626393>.

11.2 Publications of the year

International journals

- [26] X. Alameda-Pineda, A. Addelee, D. H. García, C. Reinke, S. Arias, F. Arrigoni, A. Auternaud, L. Blavette, C. Beyan, L. G. Camara, O. Cohen, A. Conti, S. Dacunha, C. Dondrup, Y. Ellinson, F. Ferro, S. Gannot, F. Gras, N. Gunson, R. Horaud, M. d’Inca, I. Kimouche, S. Lemaignan, O. Lemon, C. Liotard, L. Marchionni, M. Moradi, T. Pajdla, M. Pino, M. Polic, M. Py, A. Rado, B. Ren, E. Ricci, A.-S. Rigaud, P. Rota, M. Romeo, N. Sebe, W. Sieńska, P. Tanditnik, F. Tonini, N. Turro, T. Wintz and Y. Yu. ‘Socially Pertinent Robots in Gerontological Healthcare’. In: *International Journal of Social Robotics* (11th Nov. 2025), s12369-025-01330-6. DOI: [10.1007/s12369-025-01330-6](https://doi.org/10.1007/s12369-025-01330-6). URL: <https://inria.hal.science/hal-04737005> (cit. on p. 13).
- [27] L. Blavette, S. Dacunha, X. Alameda-Pineda, J. Cattoni, A.-S. Rigaud and M. Pino. ‘Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Two-Wave Comparative Study on Performance, Engagement, and User Perceptions’. In: *JMIR Human Factors* 12 (3rd Dec. 2025), e81936. DOI: [10.2196/81936](https://doi.org/10.2196/81936). URL: <https://inria.hal.science/hal-05469007> (cit. on p. 13).
- [28] L. Blavette, S. Dacunha, X. Alameda-Pineda, D. Hernández García, S. Gannot, F. Gras, N. Gunson, S. Lemaignan, M. Polic, P. Tanditnik, F. Tonini, A.-S. Rigaud and M. Pino. ‘Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation’. In: *JMIR Human Factors* 12 (1st Aug. 2025). DOI: [10.2196/76496](https://doi.org/10.2196/76496). URL: <https://hal.science/hal-05430362>.

- [29] G. Gomez-Trenado, P. Mesejo, O. Cordón and S. Lathuilière. ‘Don’t Forget Your Inverse DDIM for Image Editing’. In: *IEEE Computational Intelligence Magazine* 20.3 (Aug. 2025), pp. 10–18. DOI: [10.1109/MCI.2025.3563859](https://doi.org/10.1109/MCI.2025.3563859). URL: <https://hal.science/hal-05460458>.
- [30] T. de Min, M. Mancini, S. Lathuilière, S. Roy and E. Ricci. ‘Unlearning Personal Data from a Single Image’. In: *Transactions on Machine Learning Research Journal* (2026). URL: <https://hal.science/hal-05460450>. In press.
- [31] T. de Min, S. Roy, S. Lathuilière, E. Ricci and M. Mancini. ‘Group-robust Machine Unlearning’. In: *Transactions on Machine Learning Research Journal* (2026). URL: <https://hal.science/hal-05460442>. In press.
- [32] M. Sadeghi, J.-E. Ayilo, R. Serizel and X. Alameda-Pineda. ‘Posterior Transition Modeling for Unsupervised Diffusion-Based Speech Enhancement’. In: *IEEE Signal Processing Letters* 32 (2025), pp. 2694–2698. DOI: [10.1109/LSP.2025.3583967](https://doi.org/10.1109/LSP.2025.3583967). URL: <https://hal.science/hal-05135495>.
- [33] S. Sadok, S. Leglaive and R. Séguier. ‘A vector quantized masked autoencoder for audiovisual speech emotion recognition’. In: *Computer Vision and Image Understanding* 257 (June 2025), p. 104362. DOI: [10.1016/j.cviu.2025.104362](https://doi.org/10.1016/j.cviu.2025.104362). URL: <https://hal.science/hal-05041905>.

International peer-reviewed conferences

- [34] J.-E. Ayilo, M. Sadeghi, R. Serizel and X. Alameda-Pineda. ‘Diffusion-based Unsupervised Audiovisual Speech Enhancement’. In: *ICASSP 2025 - International Conference on Acoustics Speech and Signal Processing*. Hyderabad, India: IEEE, 2025, pp. 1–5. URL: <https://hal.science/hal-04718254>.
- [35] A. Belaref, S. Sadok, K. Ibrahim, Z. Noumir and R. Segulier. ‘Can AI Decode the Circumplex Model of Affect? A Data-driven Study’. In: *Pattern Recognition. ICPR 2024 International Workshops and Challenges. ICPR 2024. Lecture Notes in Computer Science, vol 15614*. Springer. International Conference on Pattern Recognition, ICPR 2024. Vol. 15614. Lecture Notes in Computer Science. Kolkata, India: Springer Nature Switzerland; Springer Nature Switzerland, 10th May 2025, pp. 97–108. DOI: [10.1007/978-3-031-87657-8_7](https://doi.org/10.1007/978-3-031-87657-8_7). URL: <https://inria.hal.science/hal-05085528>.
- [36] G. Fiche, S. Leglaive, X. Alameda-Pineda and F. Moreno-Noguer. ‘MEGA: Masked Generative Autoencoder for Human Mesh Recovery’. In: *Proc. of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR 2025 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville (Tennessee), United States: IEEE, 2025, pp. 1–16. URL: <https://hal.science/hal-04980723>.
- [37] I. E. Marouf, E. Tartaglione, S. Lathuilière and J. van de Weijer. ‘Ask and Remember: A Questions-Only Replay Strategy for Continual Visual Question Answering’. In: *ICCV 2025 - International Conference on Computer Vision*. Honolulu, United States, 19th Oct. 2025. URL: <https://hal.science/hal-05460430>.
- [38] S. Sadok, J. Hauret and E. Bavu. ‘Bringing Interpretability to Neural Audio Codecs’. In: *Interspeech 2025 - 26th edition of the Interspeech Conference*. Rotterdam, Netherlands, Aug. 2025, pp. 1–5. URL: <https://hal.science/hal-05098131>.
- [39] S. Sadok, S. Leglaive, L. Girin, G. Richard and X. Alameda-Pineda. ‘AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder’. In: *ICASSP 2025 - IEEE International Conference on Acoustics, Speech, and Signal Processing*. Hyderabad, India: IEEE, 9th Jan. 2025, pp. 1–5. URL: <https://hal.science/hal-04891286>.
- [40] Y. Zhu, X. Wang, S. Lathuilière and V. Kalogeiton. ‘Di[M]O: Distilling Masked Diffusion Models into One-step Generator’. In: *2025 International Conference on Computer Vision (ICCV 2025)*. Hawaii, United States, 19th Oct. 2025. URL: <https://hal.science/hal-05413533>.

National peer-reviewed Conferences

- [41] S. Sadok, J. Hauret and E. Bavu. ‘Donner du sens aux Codecs Neuronaux : Interprétabilité des Tokens discrets produits pour des Signaux Vocaux’. In: CFA 2025 - 17e Congrès Français d’Acoustique. Paris, France, 2025. URL: <https://hal.science/hal-05069762>.

Reports & preprints

- [42] J.-E. Ayilo, M. Sadeghi, R. Serizel and X. Alameda-Pineda. *Diffusion-based Frameworks for Unsupervised Speech Enhancement*. 15th Jan. 2026. URL: <https://hal.science/hal-05458941>.
- [43] Y. Benigmim, S. Roy, K. Oublal, I. E. Marouf, S. Essid, V. Kalogeiton and S. Lathuilière. *Make me an Expert: Distilling from Generalist Black-Box Models into Specialized Models for Semantic Segmentation*. 2025. DOI: [10.48550/arXiv.2509.00509](https://doi.org/10.48550/arXiv.2509.00509). URL: <https://hal.science/hal-05325213>.
- [44] S. Kammoun, X. Alameda-Pineda and S. Leglaive. *Modeling strategies for speech enhancement in the latent space of a neural audio codec*. 2025. URL: <https://hal.science/hal-05335192>.

11.3 Cited publications

- [45] A. Ballou, X. Alameda-Pineda and C. Reinke. *Variational Meta Reinforcement Learning for Social Robotics*. 20th Dec. 2022. URL: <https://hal.inria.fr/hal-03908505> (cit. on p. 7).
- [46] X. Lin, L. Girin and X. Alameda-Pineda. *Unsupervised Multiple-Object Tracking with a Dynamical Variational Autoencoder*. 22nd Feb. 2022. URL: <https://hal.inria.fr/hal-03584014> (cit. on p. 7).
- [47] C. Reinke and X. Alameda-Pineda. *Successor Feature Representations*. May 2022. URL: <https://hal.inria.fr/hal-03426870> (cit. on p. 6).
- [48] T. Afouras, A. Owens, J. S. Chung and A. Zisserman. ‘Self-supervised learning of audio-visual objects from video’. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 208–224 (cit. on p. 8).
- [49] S. Ba, X. Alameda-Pineda, A. Xompero and R. Horaud. ‘An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes’. In: *Computer Vision and Image Understanding* 153 (Dec. 2016), pp. 64–76. DOI: [10.1016/j.cviu.2016.07.006](https://doi.org/10.1016/j.cviu.2016.07.006). URL: <https://hal.inria.fr/hal-01349763> (cit. on p. 7).
- [50] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba and R. Horaud. ‘Tracking a Varying Number of People with a Visually-Controlled Robotic Head’. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vancouver, Canada: IEEE, Sept. 2017, pp. 4144–4151. DOI: [10.1109/IROS.2017.8206274](https://doi.org/10.1109/IROS.2017.8206274). URL: <https://hal.inria.fr/hal-01542987> (cit. on p. 7).
- [51] Y. Ban, X. Alameda-Pineda, C. Evers and R. Horaud. ‘Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM’. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 798–802. DOI: [10.1109/LSP.2019.2908376](https://doi.org/10.1109/LSP.2019.2908376). URL: <https://hal.inria.fr/hal-01969050> (cit. on p. 7).
- [52] D. Bršćić, H. Kidokoro, Y. Suehiro and T. Kanda. ‘Escaping from children’s abuse of social robots’. In: *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. 2015, pp. 59–66 (cit. on p. 10).
- [53] W.-L. Chang, J. P. White, J. Park, A. Holm and S. Šabanović. ‘The effect of group size on people’s attitudes and cooperative behaviors toward robots in interactive gameplay’. In: *RO-MAN International Symposium on Robot and Human Interactive Communication*. IEEE. 2012, pp. 845–850 (cit. on p. 9).
- [54] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson and K. Grauman. ‘Soundspaces: Audio-visual navigation in 3d environments’. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 17–36 (cit. on p. 8).

- [55] M. E. Foster, A. Gaschler and M. Giuliani. ‘Automatically classifying user engagement for dynamic multi-party human–robot interaction’. In: *International Journal of Social Robotics* 9.5 (2017), pp. 659–674 (cit. on p. 10).
- [56] R. Gao and K. Grauman. ‘Visualvoice: Audio-visual speech separation with cross-modal consistency’. In: *IEEE/CVF CVPR*. 2021 (cit. on p. 8).
- [57] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. ‘Dynamical Variational Autoencoders: A Comprehensive Review’. In: *Foundations and Trends in Machine Learning* 15.1-2 (Dec. 2021), pp. 1–175. DOI: [10.1561/22000000089](https://doi.org/10.1561/22000000089). URL: <https://inria.hal.science/hal-02926215> (cit. on pp. 6, 7).
- [58] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. ‘Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments’. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (Mar. 2019), pp. 88–103. DOI: [10.1109/JSTSP.2019.2903472](https://doi.org/10.1109/JSTSP.2019.2903472). URL: <https://hal.inria.fr/hal-01851985> (cit. on p. 7).
- [59] S. Sebo, B. Stoll, B. Scassellati and M. F. Jung. ‘Robots in groups and teams: a literature review’. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–36 (cit. on p. 9).
- [60] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 10).
- [61] M. Żarkowski. ‘Multi-party turn-taking in repeated human–robot interactions: an interdisciplinary evaluation’. In: *International Journal of Social Robotics* 11.5 (2019), pp. 693–707 (cit. on p. 10).
- [62] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker and W. Burgard. ‘Vr-goggles for robots: Real-to-sim domain adaptation for visual control’. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1148–1155 (cit. on p. 7).