Learning for Robots in Conversational Groups

Workshop ILLS – DATAIA Institute
May 25th, 2023

Xavier Alameda-Pineda (and a long list of great people)
RobotLearn Team, Inria at University Grenoble Alpes

Robots in Conversational Groups

Physical autonomous computing systems able to naturally participate in conversational groups.

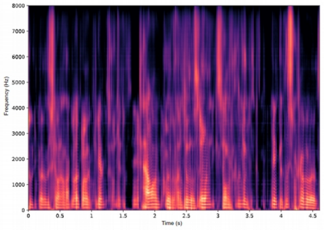
- Computer vision (person detection and tracking)
- Audio processing (speech enhancement and recognition)
- Multi-modal fusion (robust perception)
- Robot control

We tackle all this with *machine learning*.

Menu

Robot skills in populated environments

(general discussion)



Noise-agnostic speech enhancement

(dynamical VAE)

Social robot navigation

(transfer representation learning)

Robot skills in populated environments

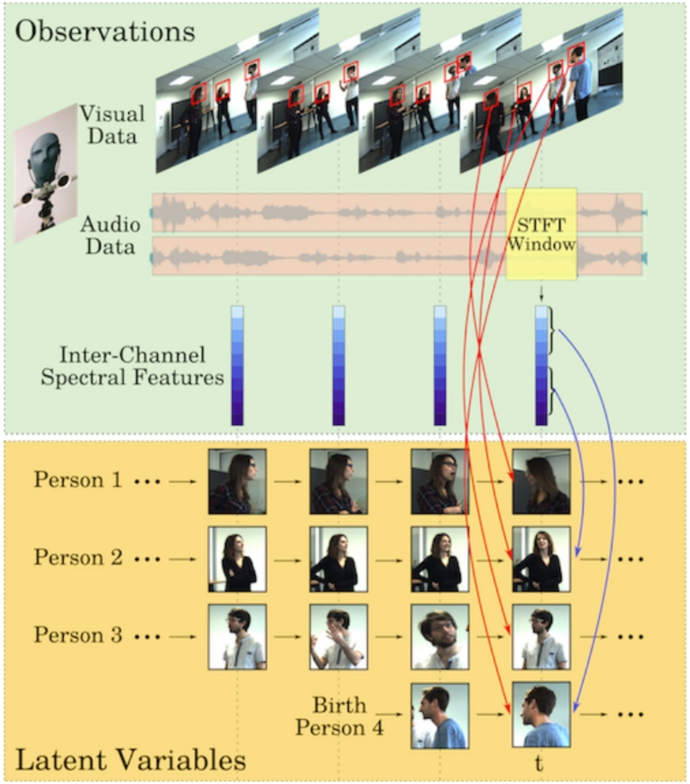
Navigate safely, participate in conversations…

- ➢ where are people/objects?
- ➢ who is speaking?
- ➢ to whom?

We have few/no access to the end-user environment beforehand → how to learn to prepare the robot for real world?

Multiple Person Tracking

In crowded scenarios[1]
and with audio-visual data.[2]



[1] Xu et al (2022).TransCenter: Transformers with Dense Representations for Multiple-Object Tracking. In IEEE TPAMI.
[2] Ban et. al. (2021). Variational bayesian inference for audio-visual tracking of multiple speakers. In IEEE TPAMI.

Generating (Robot) Behavior

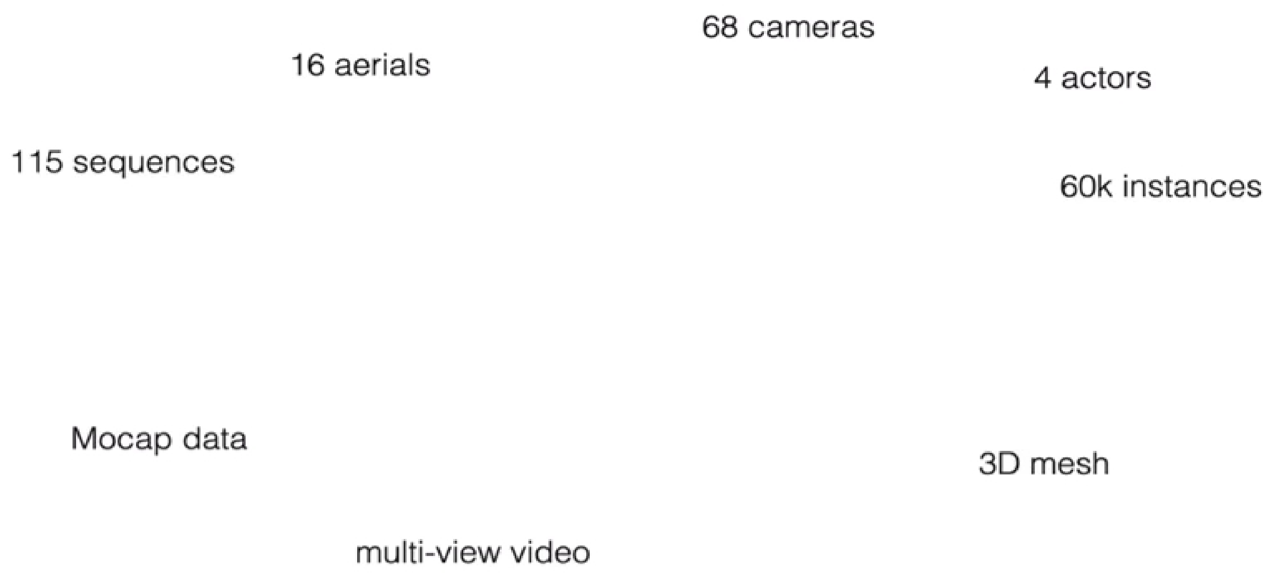
How to generate a sequence of actions **within** a conversation?

Rather generate a sequence of interleaved actions.

Also interested in co-speech gesture generation.

[3] Airale et. al. (2022). Social Interaction GAN: Multiperson interaction sequence generation. In IEEE TAFFC.

Complex Multi-person Interactions[4]



[4] Guo et. al. (2022). Multi-Person Extreme Motion Prediction. In IEEE/CVF CVPR.

Dynamical VAE for domain-free adaptation in speech enhancement[5]

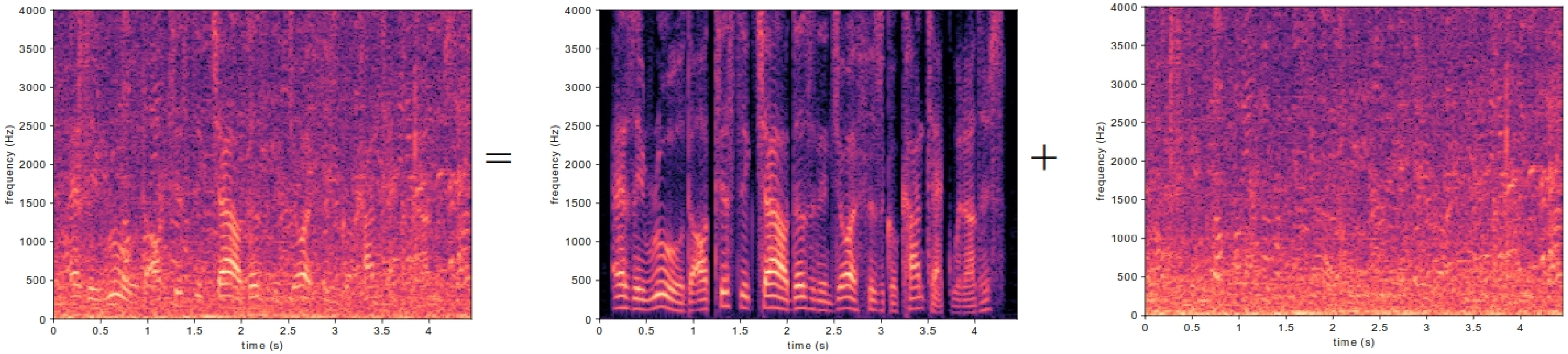
Xiaoyu Bie @ Inria
Laurent Girin @ GIPSA-Lab
Simon Leglaive @ CentraleSupélec

gipsa-lab

CentraleSupélec

[5] Bie et. al. (2022). Unsupervised speech enhancement using dynamical variational autoencoders. In IEEE TASLP.

Noise-agnostic speech enhancement?



- Speech enhancement: extracting clean speech signal from noisy recording
- Noise-agnostic: properties of noise signal **not available** at training time
  (but we do have access to clean speech)

Probabilistic modeling with VAE

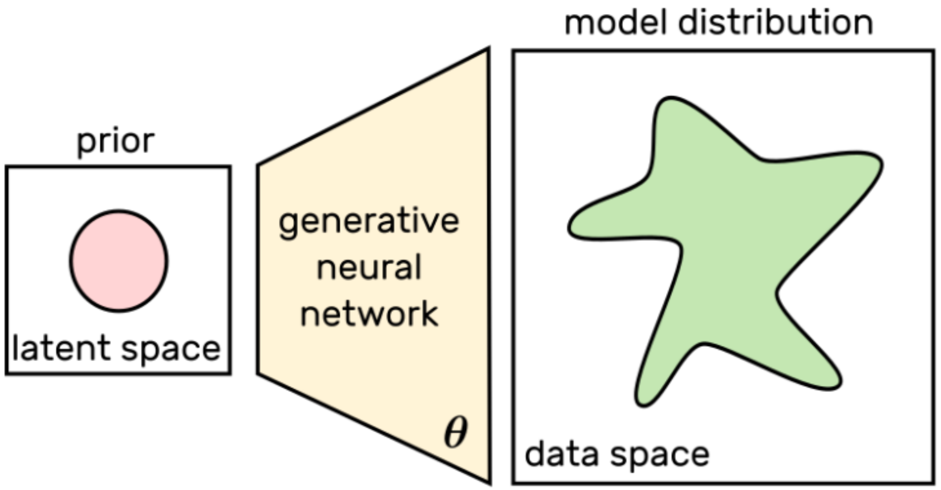
Deep latent-variable generative models can be combined with other probabilistic models at test time.

$$p(\mathbf{x};\theta) = \int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z})d\mathbf{z}.$$

Conditional likelihood modeled with the decoder network:

$$p(\mathbf{x}|\mathbf{z};\theta) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}), \mathrm{diag}\left\{\mathbf{v}_\theta(\mathbf{z})\right\}\right)$$

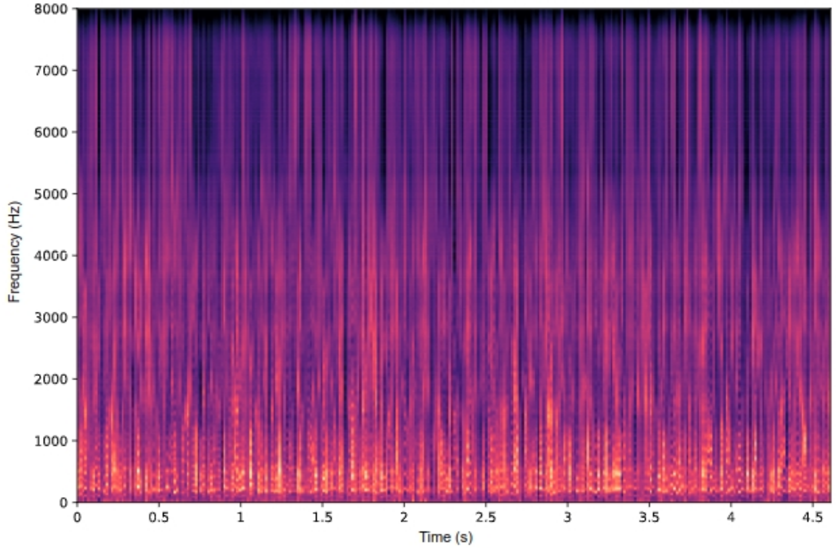
Trained by maximising the ELBO.[6]



[6] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In ICLR.

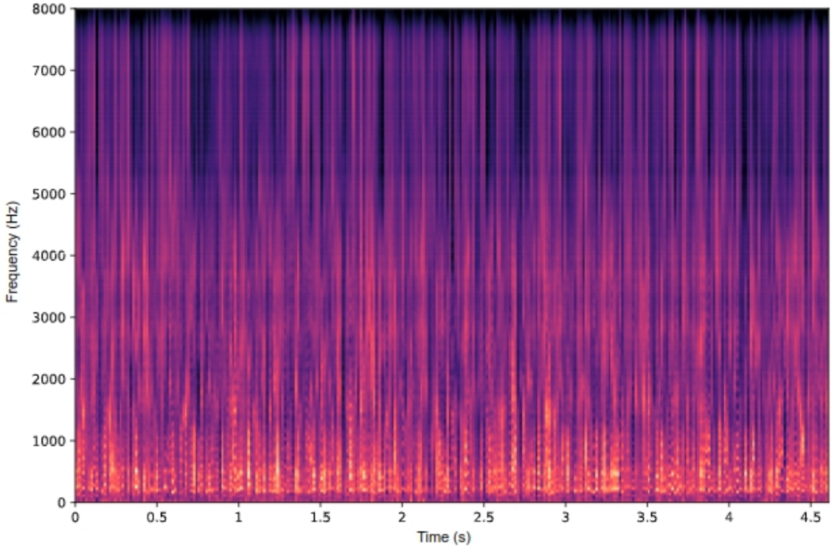
Towards Dynamical VAE

$$p_\theta^{\text{VAE}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t, \mathbf{z}_t)$$



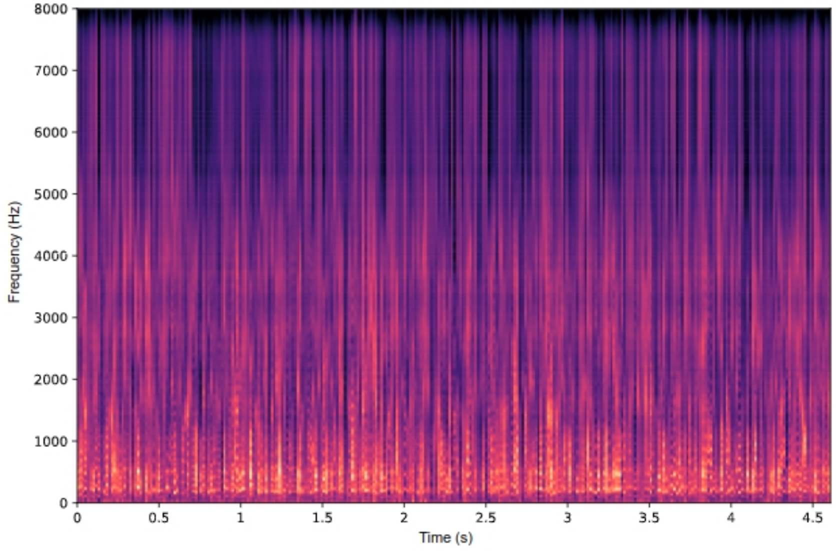
Towards Dynamical VAE

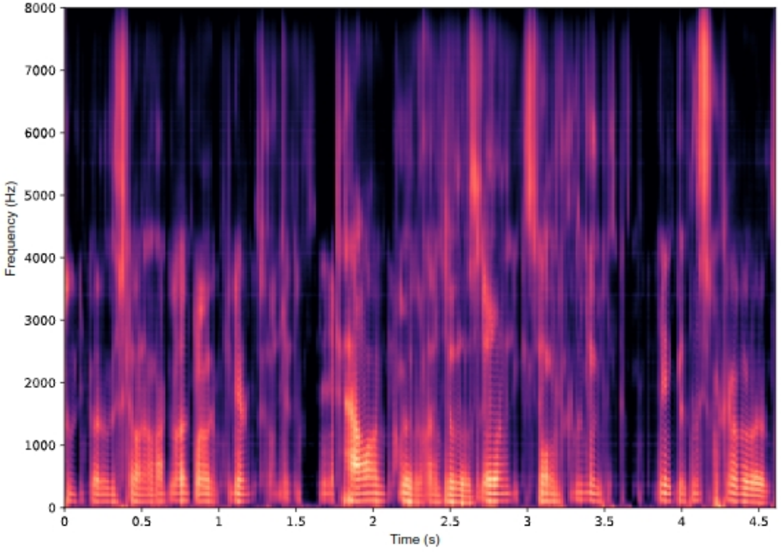
VAE treat frames independently: $p_\theta^{\text{VAE}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t, \mathbf{z}_t)$



Towards Dynamical VAE

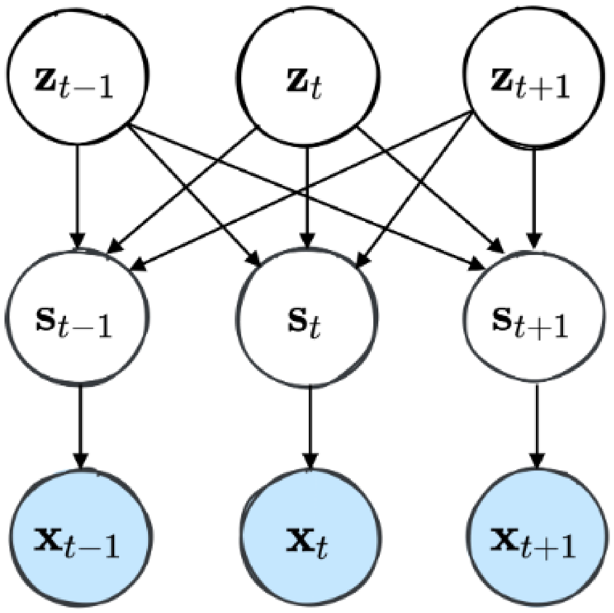
VAE treat frames independently: $p_\theta^{\text{VAE}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t, \mathbf{z}_t)$





Dynamical VAEs are able to take dependencies into account.

Back to noise-agnostic SE



Speech model (**s**'s and **z**'s) pre-trained with large dataset. Noise model for the **x**'s

The parameters of **x** need to be estimated at test time for every noisy speech recording. (EM algorithm)

DVAE:[7] family of deep probabilistic models with temporal dependencies

[7] Girin, L et al (2021). Dynamical variational autoencoders: A comprehensive review. FnT Machine Learning.

Results & Discussion

Method *vs.* Setting	D1→D1	D2→D2
Supervised	5.7	14.0
Noise dependent	–	17.7
Noise agnostic (ours)	5.8	17.1

Results & Discussion

Method *vs.* Setting	D1→D1	D2→D2	D2→D1	D1→D2
Supervised	5.7	14.0	4.1	10.4
Noise dependent	–	17.7	-1.6	–
Noise agnostic (ours)	5.8	17.1	4.6	17.3

Results & Discussion

Method *vs.* Setting	D1→D1	D2→D2	D2→D1	D1→D2
Supervised	5.7	14.0	4.1	10.4
Noise dependent	–	17.7	-1.6	–
Noise agnostic (ours)	5.8	17.1	4.6	17.3

- SI-SDR on standard datasets (other metrics available in the paper)
- Hear examples at: https://team.inria.fr/robotlearn/unsup-se-dvae/

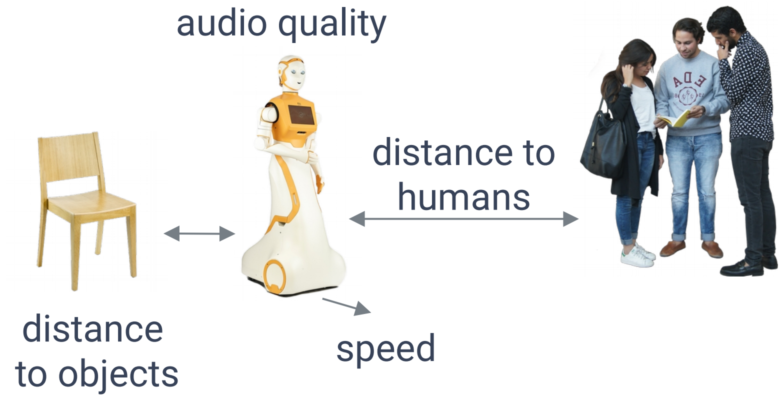
SFR: Successor Feature Representations

Chris Reinke @ Inria

[8] Reinke, C et al (2023). Successor Feature Representations, Accepted at TMLR.

Goal

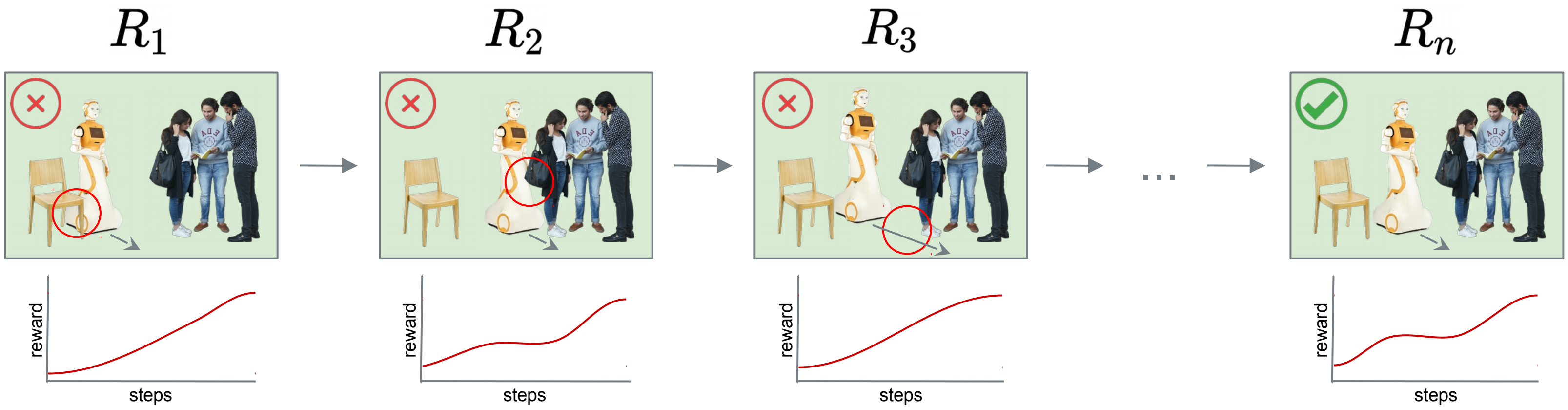
Humanoid robot to join a group of people → Deep RL?
Good behavior depends on many components



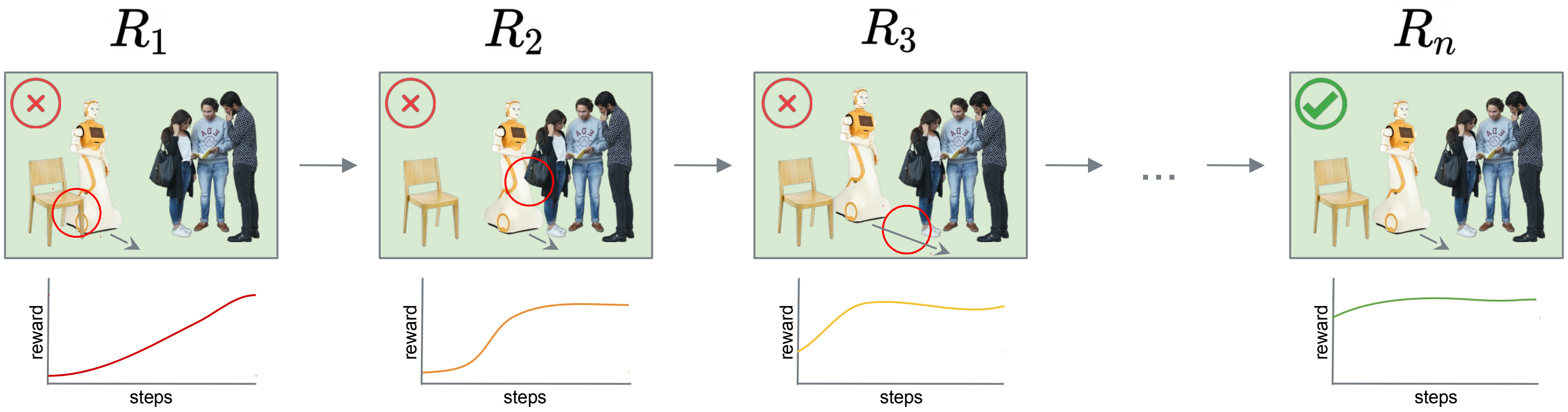
At least two issues:

- ➢ Data (lots needed, but expensive)
- ➢ Which combination of reward components results in desired behaviors?

How to find a good Reward Function?

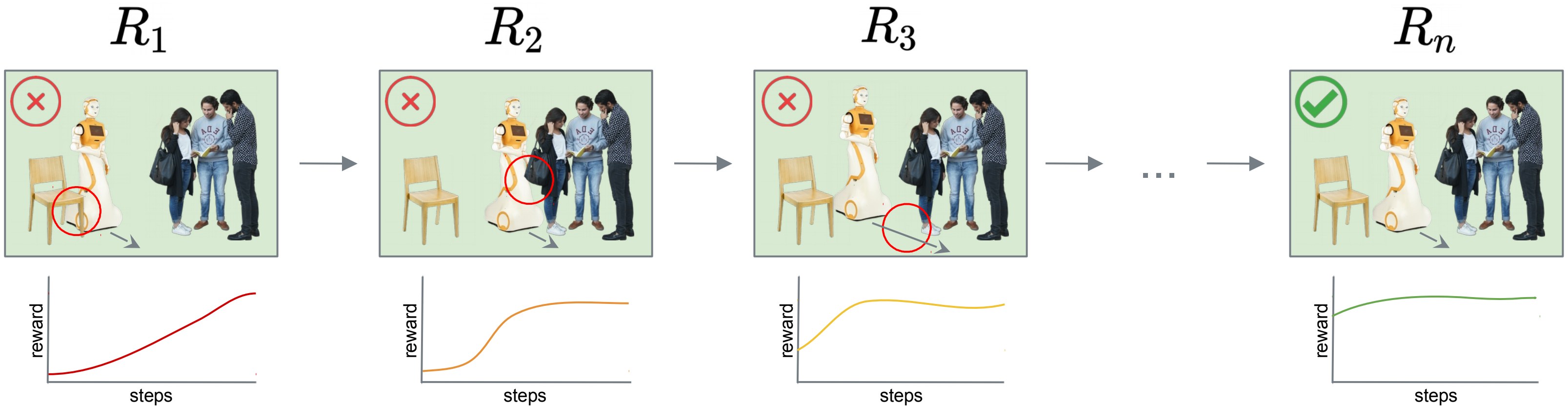


How to find a good Reward Function?



Main idea: accumulate knowledge over tasks (reward combinations):

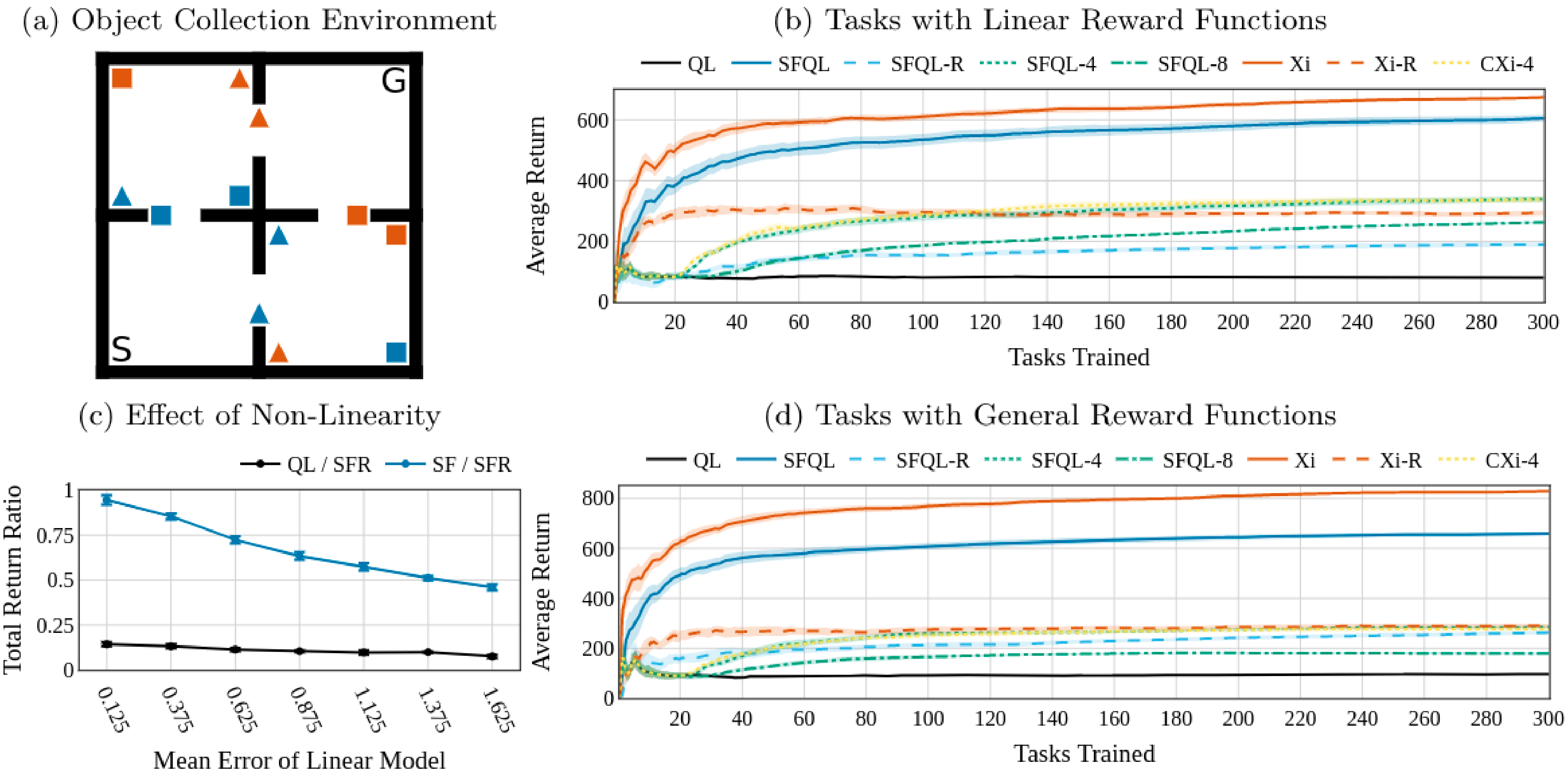
How to find a good Reward Function?



Main idea: accumulate knowledge over tasks (reward combinations):

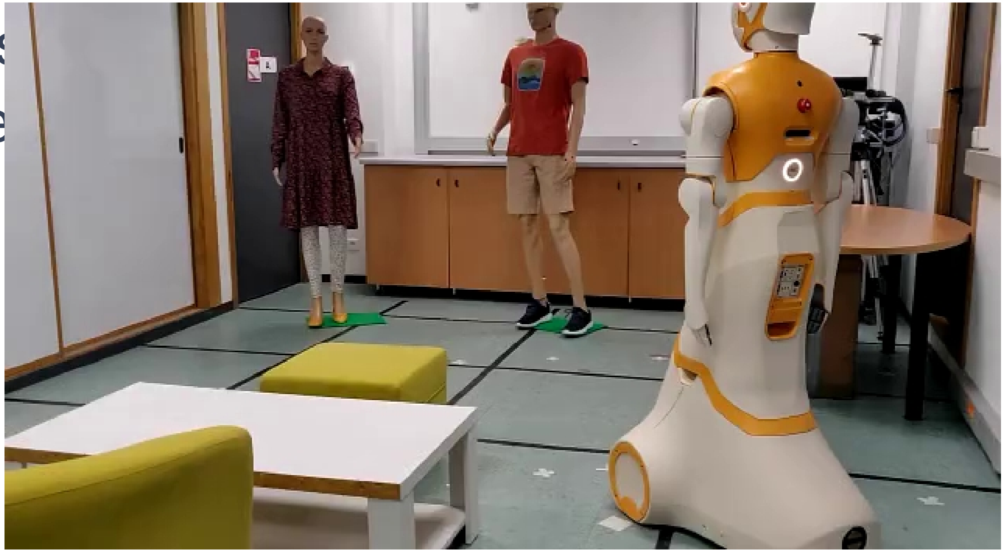
- Previous model: linear combinations. $r(s, a, s') = \boldsymbol{\phi}(s, a, s')^\top \mathbf{w}$
- Our proposal: non-linear (arbitrary). $r(s_t, a_t, s_{t+1}) \equiv R(\phi(s_t, a_t, s_{t+1})) = R(\phi_t)$
- We proposed a new Bellman/learning operator, proved convergence, and all theoretical and algorithmic developments.

SFR – Results


(a) Object Collection Environment
G
S
(b) Tasks with Linear Reward Functions
QL
SFQL
SFQL-R
SFQL-4
SFQL-8
Xi
Xi-R
CXi-4
Average Return
0
200
400
600
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Tasks Trained
(c) Effect of Non-Linearity
QL / SFR
SF / SFR
Total Return Ratio
0
0.25
0.5
0.75
1
0.125
0.375
0.625
0.875
1.125
1.375
1.625
Mean Error of Linear Model
(d) Tasks with General Reward Functions
QL
SFQL
SFQL-R
SFQL-4
SFQL-8
Xi
Xi-R
CXi-4
Average Return
0
200
400
600
800
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Tasks Trained

Conclusion

- Working with social robots requires robustness to environment change



- This cannot come at the cost of requiring large annotated datasets
- We need to deal with the lack of annotated data/environment interactions
- Unsupervised DA, meta/transfer learning, are possible ways, but they need to be tested in real conditions.

Thanks

All **RobotLearn**-ers and collaborators!

The funding programs supporting our research:

SPRING

MIAI Grenoble Alpes

ANR — PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

And you for listening. Question (and answer?) time!

→ Available for discussion: let's mingle!