ANITI
RECHERCHE

Funded by
the European Union

erc
European Research Council
Established by the European Commission

C3PO =
Language + Vision + Robotics

Rufin VanRullen

cnrs

CerCo
UMR5549

NeuroAi

Synergy Chair C3PO (2024-2028)

ANITI

RECHERCHE

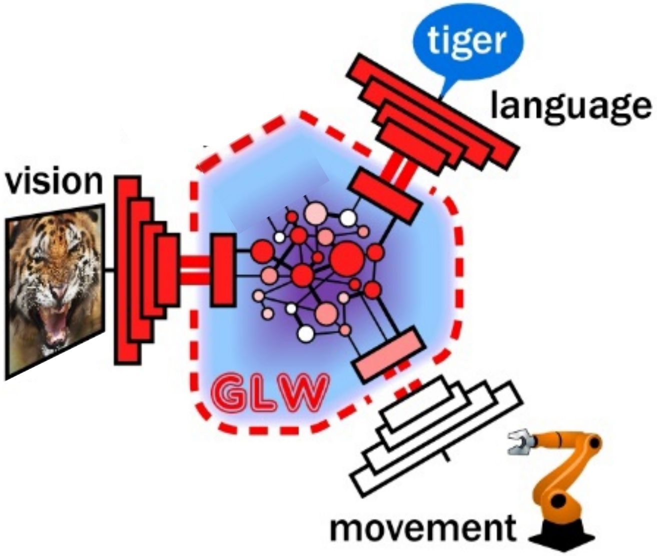
C3PO = Cobots with Conversation, Cognition & Perception

4 Chairs:

- R.VanRullen (CerCo) – Brain-inspired Deep Learning
- N. Asher (IRIT) – Linguistics
- T. Serre (Brown) – Vision
- O. Stasse (LAAS) – Robotics

C3PO

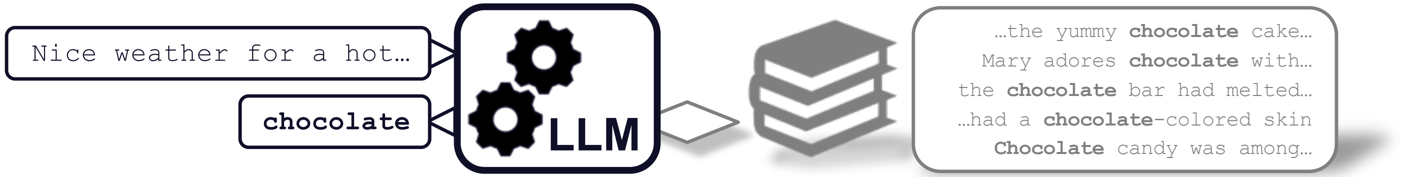
➔ Frugal multimodal robotic systems with grounded perception, language & action



Distributional vs. Referential Semantics

How do LLMs « understand »?

- **Distributional semantics: symbol meaning derived from the distribution of symbol (co-)occurrences in natural language**



Nice weather for a hot…

chocolate

…the yummy **chocolate** cake…

Mary adores **chocolate** with…

the **chocolate** bar had melted…

…had a **chocolate**-colored skin

Chocolate candy was among…

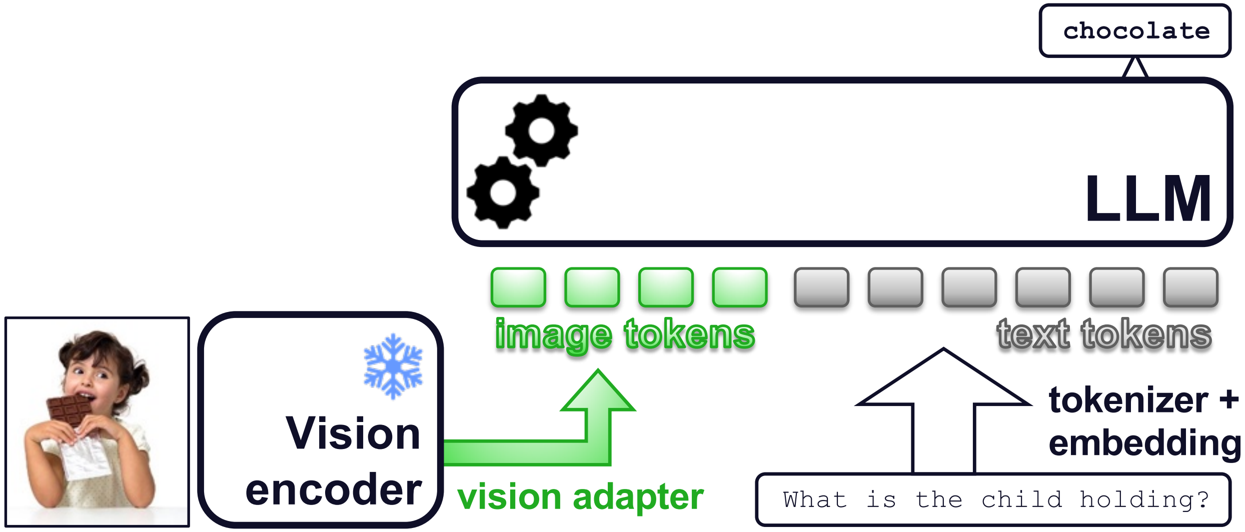
How do humans understand ?

- **Referential semantics: symbol meaning derived from its associations with other modalities (vision, touch, sensorimotor, memory, etc.) = grounding**

chocolate

Large Multimodal Models (LMMs) & Grounding

- ◎ **Recent models augment LLMs with new modalities (Vision, Action…)**
- ◎ **Does it constitute Grounding? Do LMMs have referential semantics?**



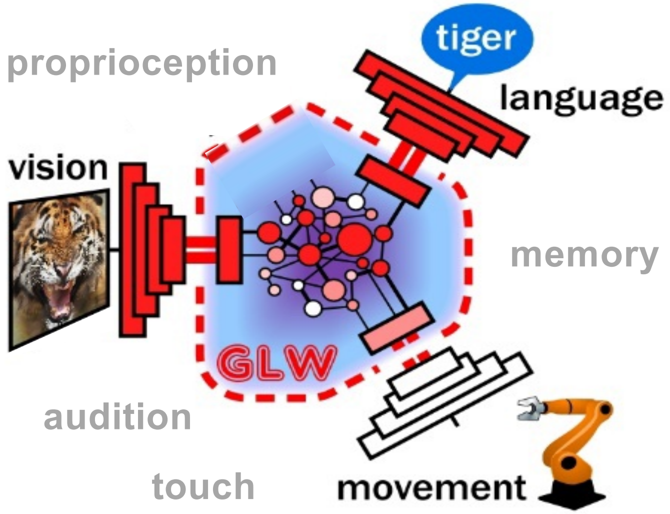
- ◎ **<u>No « real » grounding</u>: you cannot build grounding on language – language must be built on grounding!**

Rethinking grounding & language models

- LLMs use distributional semantics because it works (until it doesn't)

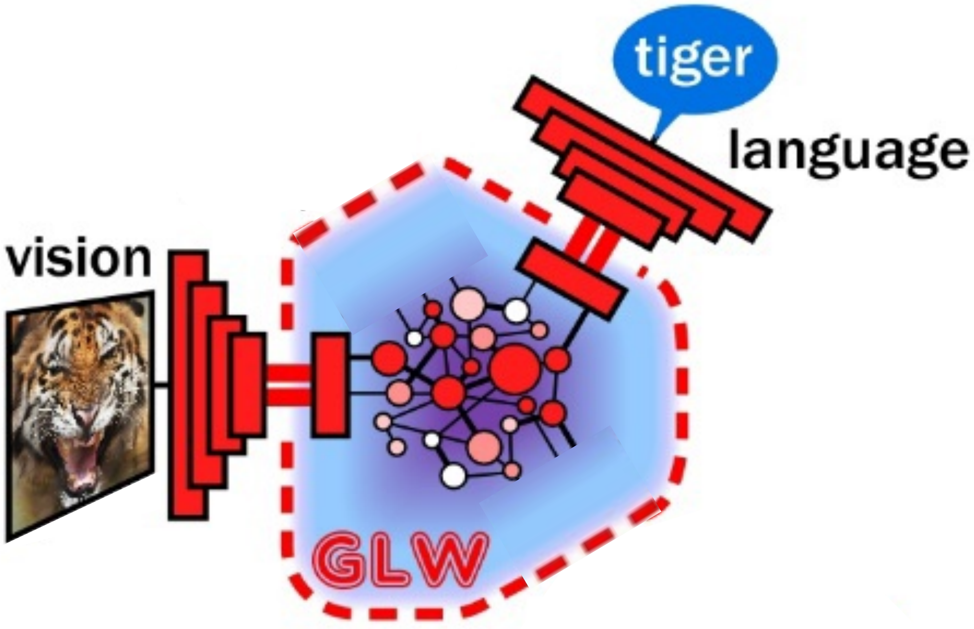
➔ We can build language models with referential semantics

- GLW = Global Latent Workspace
  - Inspired by the *Global Workspace Theory* of the brain
  - Common representation space that learns the associations or « analogies » between domains ➔ grounding, affordance
  - The GLW representation can be converted back to each input domain ➔ broadcast



Global Workspace: proof of concept

1. Vision-language grounding



Global Workspace: proof of concept

Very ambitious

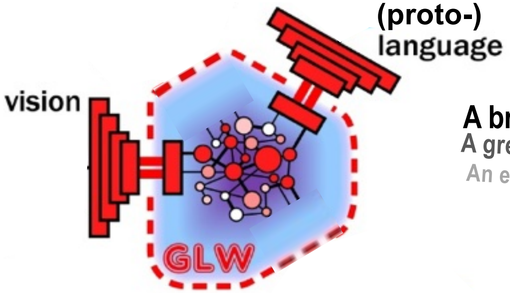
Modest

A hammock under a palm tree in a paradisiac beach scene
Children playing soccer under the rain
A man is walking his dog on a busy street in New York in front of the subway
Airport scene with a refueling truck in the foreground
Donald Trump eating a burger from McDonalds

R. Bertin-Johannet

A red chair on the left, a blue chair on the right
An orange cone in front of a blue chair
Three crates are stacked at the back of the room, next to a cone and a table

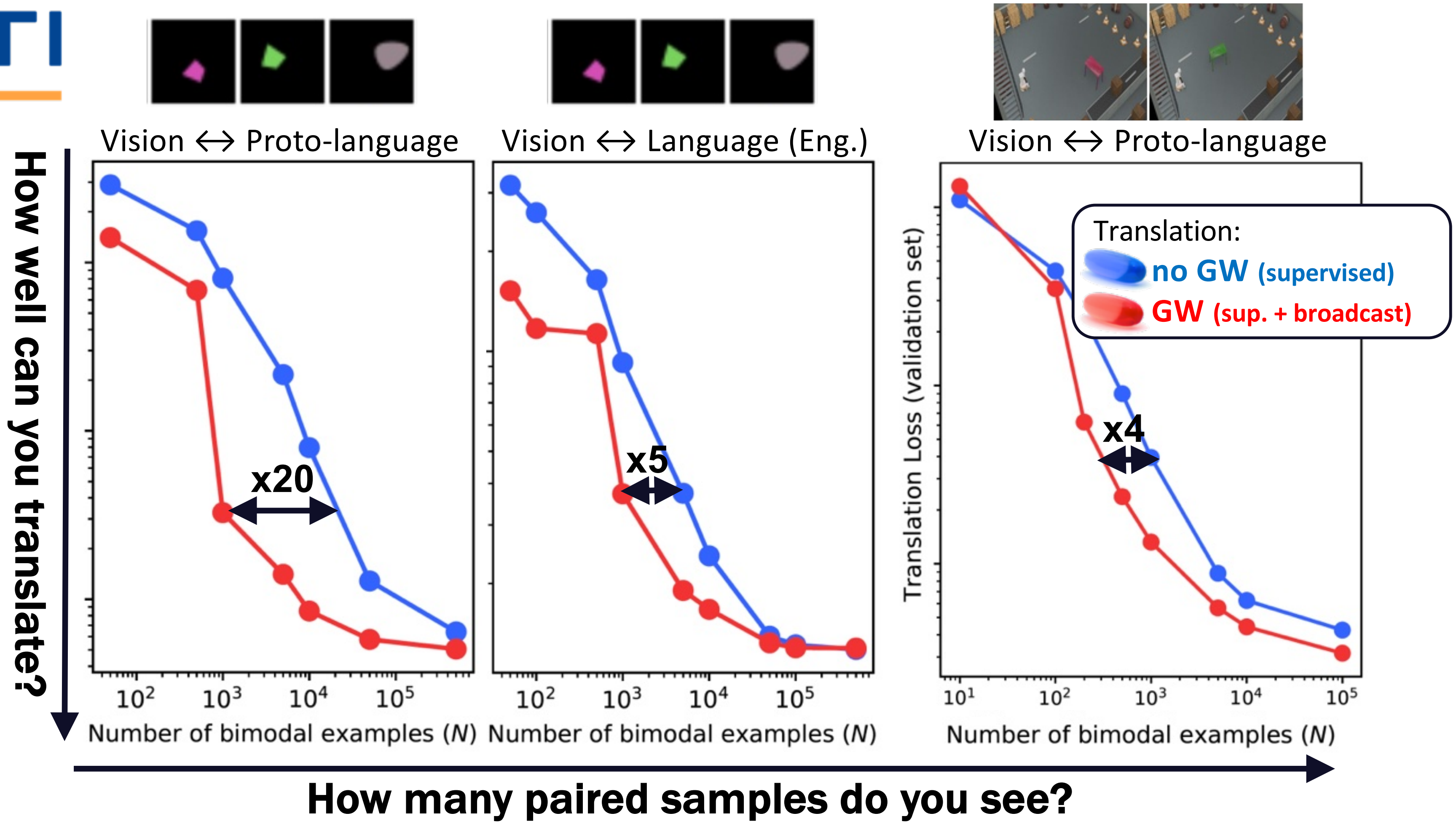
L. Maytié



A bright red diamond, pointing to the bottom right
A green triangle at the top, towards left
An egg-shaped oval, dark-blue, pointing up, on the bottom-left of the image

B. Devillers

Translation with/without GW



How well can you translate?

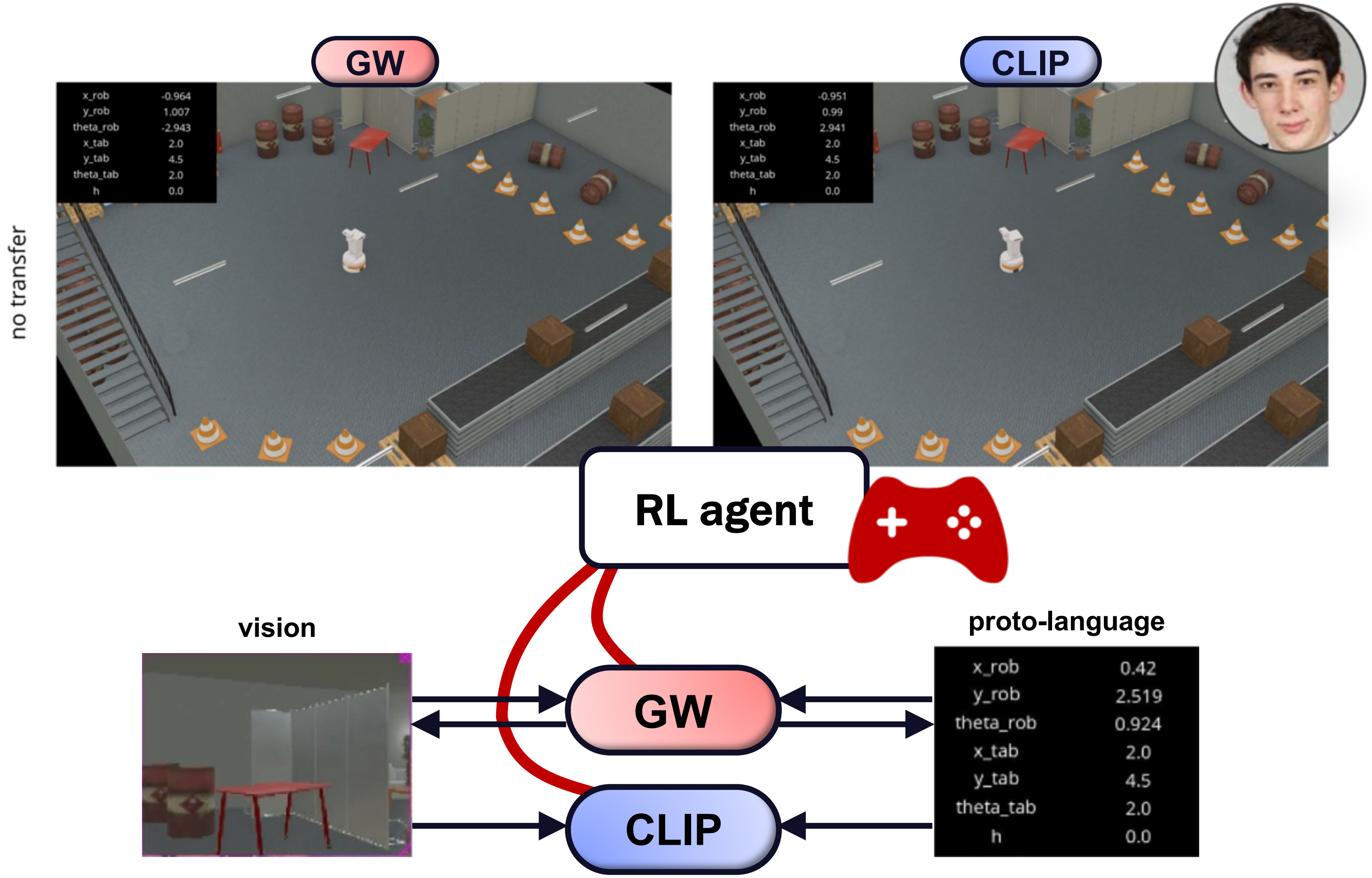
How many paired samples do you see?

➔ Image-to-Text & Text-to-image translation (DALL-E3) with 10x less supervision??

Leveraging GW-grounded representations in RL

ANITI

RECHERCHE



x_rob	-0.964
y_rob	1.007
theta_rob	-2.943
x_tab	2.0
y_tab	4.5
theta_tab	2.0
h	0.0

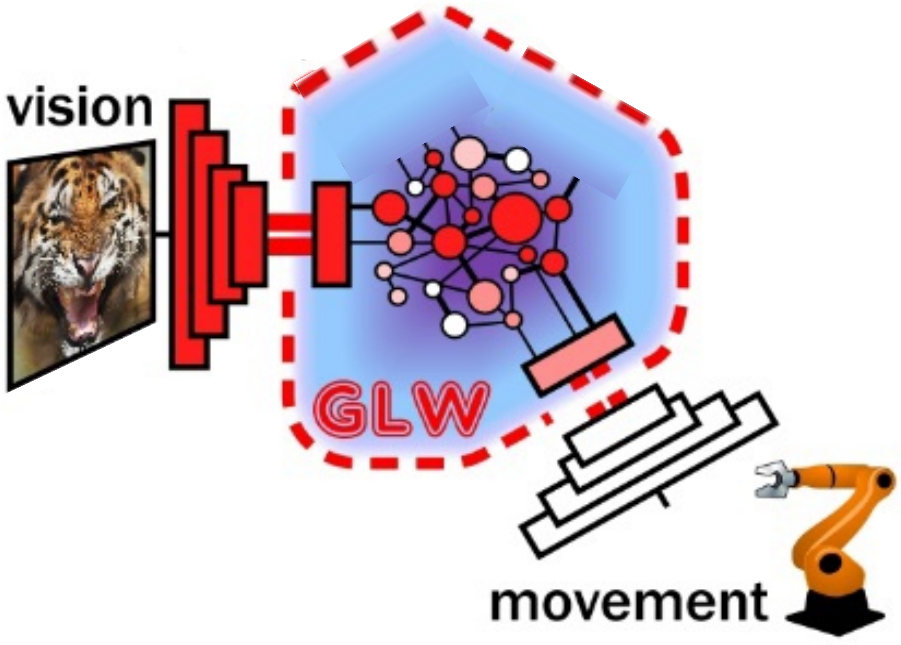
x_rob	-0.951
y_rob	0.99
theta_rob	2.941
x_tab	2.0
y_tab	4.5
theta_tab	2.0
h	0.0

x_rob	0.42
y_rob	2.519
theta_rob	0.924
x_tab	2.0
y_tab	4.5
theta_tab	2.0
h	0.0

Maytie et al, RL conference 2024

Global Workspace: proof of concept

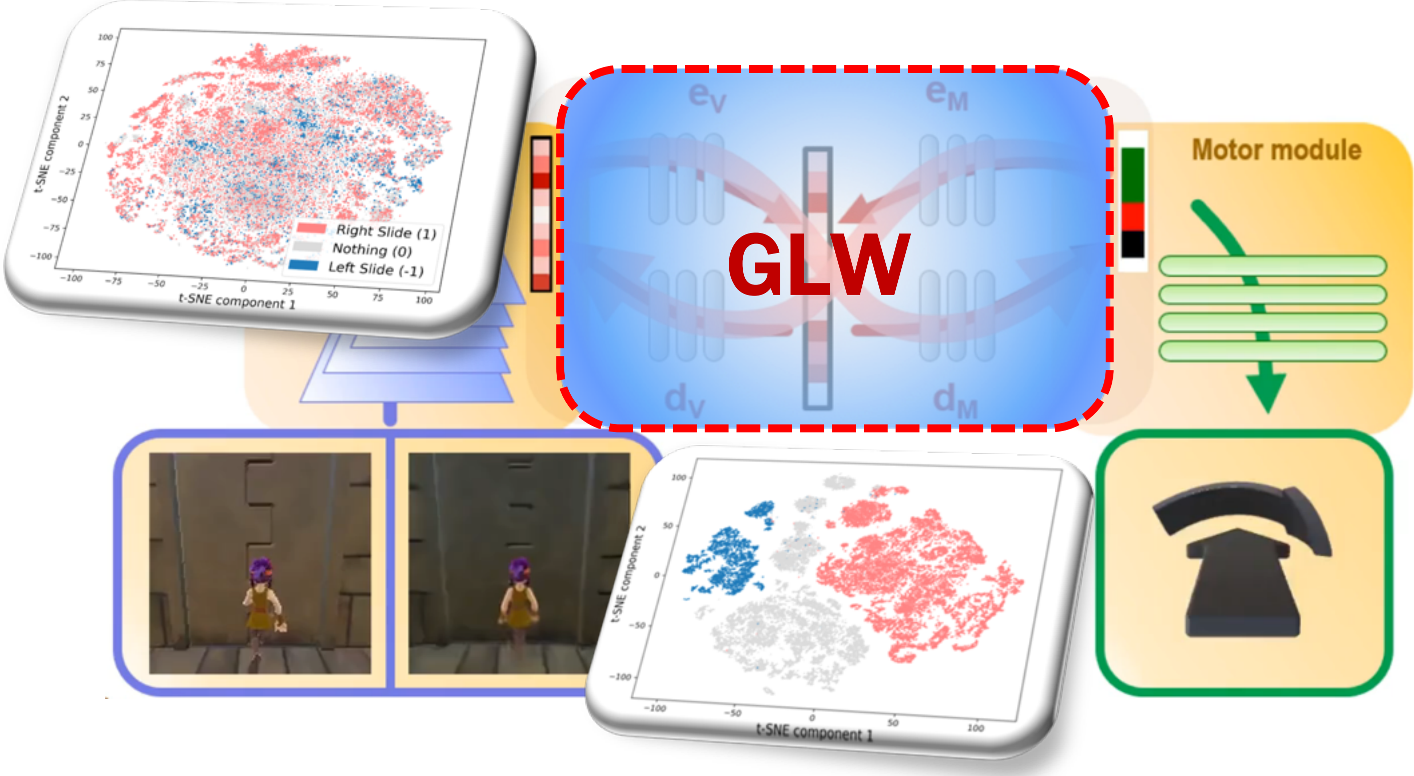
2. Vision-action grounding = *affordance*



Sensorimotor affordances in the GW

N. Kuske

- « Obstacle tower » environment
- GW visuo-motor associations learned from a pre-trained RL agent
- GW can translate & back-translate between vision and action
- The GW latent space (but not the visual one) is organized w.r.t affordances



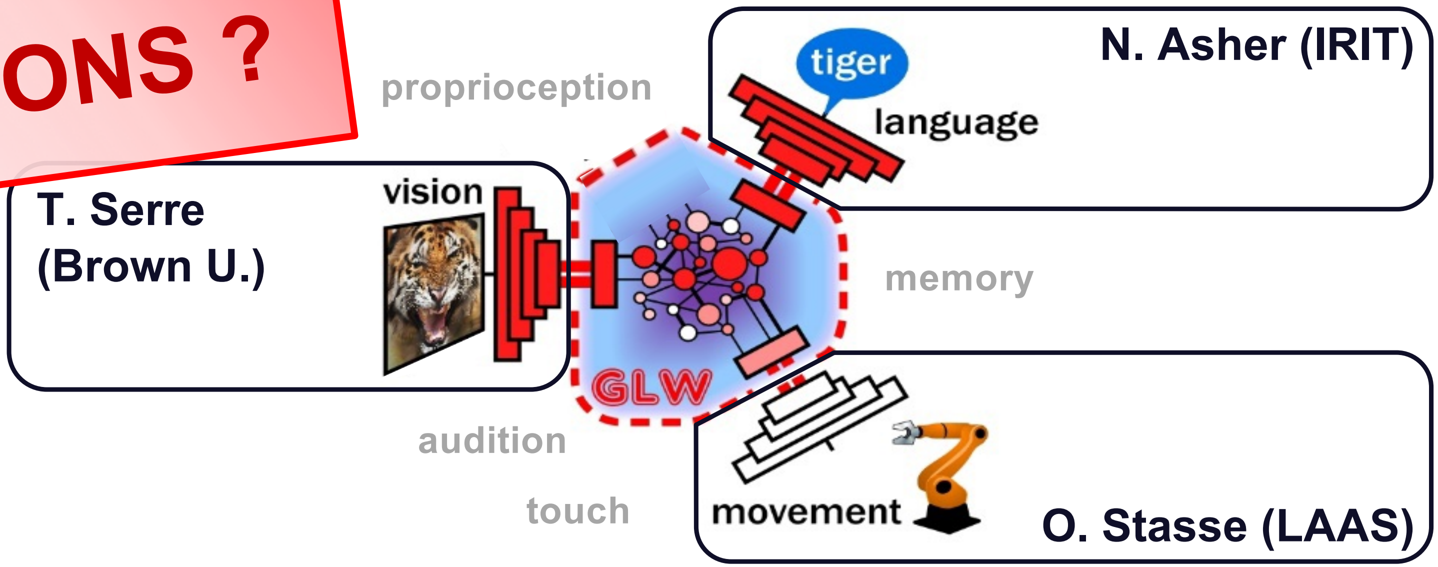
Rethinking grounding & language models

ANITI
RECHERCHE

- Using the GW framework, we can build efficient multimodal representations with grounding and affordance
- This is the first step in building LLMs (and more generally, Foundation Models) based on referential semantics

➔C3PO Synergy Chair

QUESTIONS ?



N. Asher (IRIT)
tiger
language
proprioception
T. Serre
(Brown U.)
vision
memory
GLW
audition
touch
movement
O. Stasse (LAAS)