Sensorimotor Affordances in a Global Latent Workspace

Nicolas Kuske (nicolas.kuske@gmail.com)
Rufin VanRullen (rufin.vanrullen@cnrs.fr)

CerCo, CNRS UMR5549
Artificial and Natural Intelligence Toulouse Institute
Université de Toulouse

Abstract

Understanding the role of embodiment in cognition is critical for advancing both neuroscience and artificial intelligence. While biological systems rely on multimodal sensorimotor interactions to ground meaning, artificial models often lack this grounding, limiting their ability to generalize across tasks and environments. In this work, we investigate the emergence of sensorimotor affordances within a Global Latent Workspace (GLW)—a multimodal deep learning architecture inspired by the Global Workspace Theory of consciousness. We train a reinforcement learning agent to perform a simulated embodied task (Obstacle Tower Challenge), and use its sensory-motor data to train a GLW multimodal representation (based on an encoder-decoder structure linked with each modality). We compare the GLW representation of images (from the agent's point of view) with the same image representations from a variational autoencoder. Our analysis reveals that the sensorimotor GLW compresses visual information into a structured motor latent manifold, naturally clustering affordance-relevant representations. Notably, these affordances enable zero-shot visual scene generation based on motor states, providing preliminary empirical support for sensorimotor theories of consciousness. By embedding affordances in a shared latent space, the GLW framework offers a biologically inspired path toward more generalizable and grounded artificial perception.

Keywords: embodied cognition; affordance; reinforcement learning; multimodal learning; global workspace theory; sensorimotor contingency theory; consciousness

Introduction

Sensorimotor experiences are crucial in shaping human intelligence. Motor activity from early development sculpts neural circuits for spatial and causal reasoning (Hadders-Algra, 2018). Current AI systems based mainly on language statistics lack understanding of spatiotemporal relationships (Bender & Koller, 2020; Huang et al., 2025). Although vision-language-action models offer promise, fully integrating language with sensorimotor capabilities remains a challenge, fueling interest in approaches like the embodied Turing test (Zador et al., 2023). Drawing on Global Workspace Theory (Baars, 1993; Mashour et al., 2020) and Sensorimotor Contingency Theory (O'Regan & Noë, 2001; O'Regan, 2011), we suggest that early consciousness evolved for real-time sensorimotor control. The recently proposed Global Latent Workspace (GLW) framework integrates specialized pre-trained modules using encoder-decoder pairs (Devillers et al., 2024; VanRullen & Kanai, 2021). Inspired by Gibson's affordances and studies on embodied agents (Clay et al., 2021, 2023; Gibson, 1979) our Sensorimotor GLW is trained on curated stimulus-action pairs from a high-performing RL agent and benchmarked against a Variational Autoencoder, marking a preliminary step toward integrating conscious-like processing with modern AI.

Methods

Data Collection and Environment

Sensorimotor data is gathered from an RL agent trained in the procedurally generated 3D jump-and-run puzzle Obstacle Tower Challenge environment (Juliani et al., 2019). The agent processes raw images (3×168×168) and 11-dimensional vectors representing scene and environment meta-information, outputting a 54-dimensional one-hot action vector. After 10 million training steps—when the agent typically reaches level 4—we performed inference runs collecting 850,000 image-action pairs. A video of the final, inference RL agent interacting with the environment is available at `https://youtu.be/cNMItuHy_J8`.

Unimodal Modules and Global Latent Workspace (GLW) training

The vision module compresses ≈100,000-dimensional images into 64-dimensional latent codes using a VAE trained on the above-collected vision data over 100 epochs. In contrast, the motor module does not need training. It encodes the 54-dimensional motor outputs into 4 discrete dimensions that represent forward-backward movement, left-right, jumping, and rotation. Together, these modules provide the input data for subsequent global latent workspace (GLW) training.

The GLW integrates the data from both modules into a shared latent space, ensuring alignment between modalities and enabling both input retrieval and cross-modal translation or broadcast. It consists of multiple 3-layer perceptron encoder-decoder pairs connecting each unimodal module to the central workspace as illustrated in Figure 1 (A). The encoder-decoder structures are optimized with translation, contrastive, and cycle-consistency (full and half-cycle) losses (Devillers et al., 2024). We trained 4 GLW architectures—with variations in layer sizes and loss weights—on the 0.8 million image-action pairs for 100 epochs. The subsequent analysis focuses on the best-performing architecture with 96 neurons for each coder layer, a workspace size of 32 and all loss terms having equal coefficients, except translation that is increased by a factor of 2. Other network parameters yield qualitatively similar results.

Results

Affordance Accuracy and Visual Reconstruction

Motor affordance accuracy was measured using the vision module as input and then comparing the continuous GLW motor latent outputs to the discrete RL action values as groundtruth in Figure 1 (B). Nearest neighbor mapping of GLW motor latents to discrete actions (e.g., 0.6 is mapped to 1, -0.3 to 0, and -0.7 to -1) yields 74% of correct sensorimotor affordance pairings. Visual reconstruction loss, assessed via the mean squared error comparing the original VAE latents with the latents after a complete GLW cycle (from vision to action and back), varied by action category—for instance, less frequent scenes affording jumping actions showed higher loss (range from MSE=0.4 frequent to MSE=0.7 rare event).

Sensorimotor Affordance Clusters

Using t-SNE to visualize latent representations, we observed distinct sensorimotor affordance clusters in the GLW compared to the VAE. Clusters were pronounced for forward/backward, left/right slide, and jumping actions, while rotation showed less distinct grouping. The results are illustrated for the first two motor dimensions in Figure 1 (B).

Affordance Scene Generation

The GLW's broadcasting ability enables the direct translation of motor information into visual scenes. Although decoding of the original discrete RL motor outputs as GLW inputs produced blurry images, slight deviations in the 4-dimensional motor latent space generated rich, affordance-specific scenes (e.g., forward movement produced deeper visual scenes, while walls emerged during backward movement). Videos demonstrating latent space traversal are available at `https://youtu.be/bsiUWs-81T0` (forward to backward) and `https://youtu.be/ZAls1Ws5wzU` (right to left slide).

Discussion and Outlook

State-of-the-art multimodal LLMs lack robust spatial understanding which may require embodiment (Bender & Koller, 2020; Huang et al., 2025). Our work unifies Global Workspace Theory and Sensorimotor Contingency Theory in the Global Latent Workspace (GLW), offering a mathematically tractable model that deepens our understanding of embodied intelligence and its relation to conscious-like information processing. Despite these advances, our GLW is still an encoder-decoder structure rather than a fully integrated agent. Bridging this gap by combining RL and GLW training represents a promising direction toward achieving state-of-the-art embodied robotic systems and approaching an embodied Turing test (Zador et al., 2023). Moreover, our generative results hint at the potential benefits of incorporating predictive processing with motor affordances to yield more robust, noise-resistant representations (Seth, 2014; Taniguchi et al., 2023).

Future work will focus on evaluating GLW affordances on downstream visual tasks—such as predicting distances to obstacles—to further demonstrate its practical utility. Preliminary findings suggest that GLW representations encode spatial relationships more effectively than conventional VAEs, paving the way for deeper exploration into the cognitive benefits of embodied intelligence.

Acknowledgments

Funded by the European Union (ERC Advanced GLOW project number 101096017). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

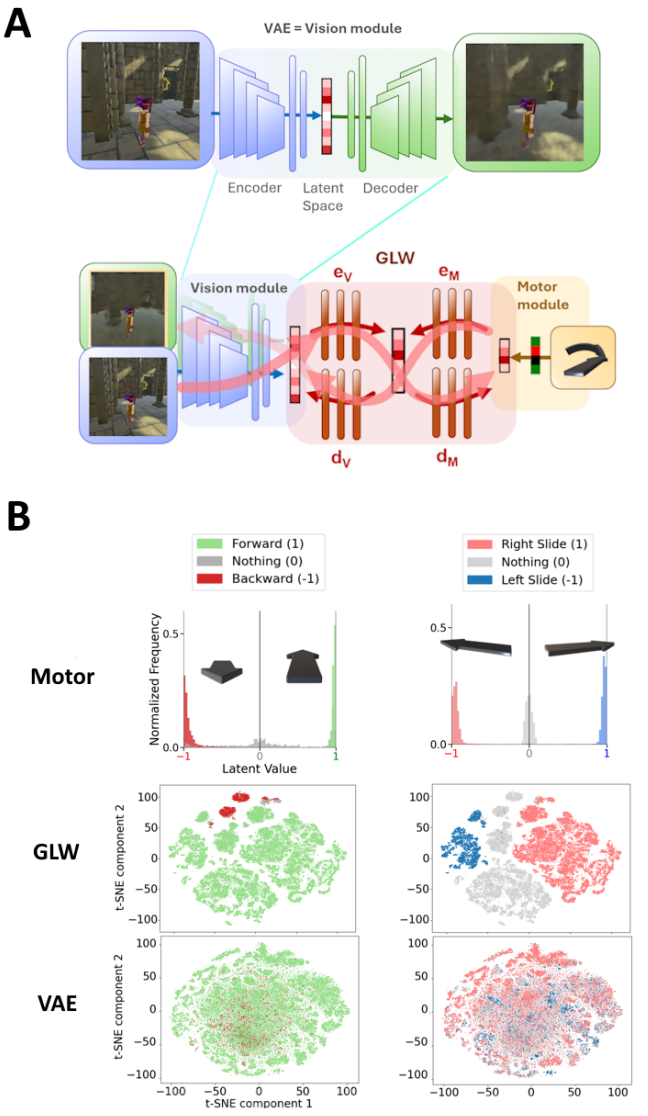

Figure 1: **A** A variational autoencoder (VAE) is trained to reconstruct the images collected during the inference runs of a reinforcement learning agent. The latent vectors from the VAE as vision module and the motor module serve as unimodal inputs to the GLW. The vision and motor encoders e_V and e_M serve to map the inputs onto a shared latent representation (the GLW proper). Decoder networks d_V and d_M serve to map the GLW representation back into the format of each unimodal's domain, thereby implementing the "broadcast" idea at the heart of GWT. **B Motor** Distribution of the GLW motor decoder d_M output for the forward-backward and left-right action dimensions. Motor latents are colored according to the discrete RL ground-truth values. E.g., green are latent motor output values of input image scenes which the RL agent associated with a forward action value of 1. **GLW** t-SNE embeddings of latent representations from the central GLW layer. Color coding as for Motor. **VAE** t-sne for VAE latents.

References

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198).

Clay, V., König, P., Kühnberger, K.-U., & Pipa, G. (2021). Learning sparse and meaningful representations through embodiment. *Neural Networks*, *134*, 23–41.

Clay, V., König, P., Kühnberger, K.-U., & Pipa, G. (2023). Development of few-shot learning capabilities in artificial neural networks when learning through self-supervised mechanisms. *Neural Networks*.

Devillers, B., Maytié, L., & VanRullen, R. (2024). Semi-supervised multimodal representation learning through a global workspace. *arXiv preprint arXiv:2306.15711*.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Hadders-Algra, M. (2018). Early human motor development: From variation to the ability to vary and adapt. *Neuroscience; Biobehavioral Reviews*, *90*, 411–427.

Huang, J.-T., Dai, D., Huang, J.-Y., Yuan, Y., Liu, X., Wang, W., ... Zhang, L. (2025). Visfactor: Benchmarking fundamental visual cognition in multimodal large language models. *arXiv preprint arXiv:2502.16435*.

Juliani, A., Berges, V.-P., Teng, E., Cohen, A., Harper, J., Elion, C., ... Lange, D. (2019). The obstacle tower: A generalization challenge in vision, control, and planning. *arXiv preprint arXiv:1902.01378*.

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798.

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97–118.

Taniguchi, T., Murata, S., Suzuki, M., Ognibene, D., Lanillos, P., Ugur, E., ... Pezzulo, G. (2023). World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges. *Advanced Robotics*, *37*(13), 780–806.

VanRullen, R., & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences*, *44*(9), 692–704.

Wu, A., Brantley, K., & Artzi, Y. (2024). A surprising failure? multimodal llms and the nlvr challenge. *arXiv preprint arXiv:2402.17793*.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., ... Tsao, D. (2023). Catalyzing next-generation artificial intelligence through neuroai. *Nature Communications*, *14*, 1597.