Modelling the Effect of Audience Tuning on Generative Episodic Memory

Zahra Fayyaz* (zahra.fayyaz@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Aya Altamimi* (aya.altamimi@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Ullrich Wagner (ullrich.wagner@uni-muenster.de)
Department of Psychology, Fliednerstr. 21
48149 Münster, Germany

Sen Cheng (sen.cheng@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Gerald Echterhoff (g.echterhoff@uni-muenster.de)
Department of Psychology, Fliednerstr. 21
48149 Münster, Germany

Laurenz Wiskott (laurenz.wiskott@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Abstract

Episodic memory is highly malleable and shaped by social interactions. Rather than storing exact representations, it reconstructs incomplete traces using semantic information influenced by motivation and context. In shared reality situations, as shown by the saying-is-believing (SIB) paradigm, verbalizing information affects both audience perception and the speaker's memory. Yet, most computational models assume faithful storage and neglect social influence. We have developed a generative episodic memory model combining a VQ-VAE for visual perception, a masking module for attention and hippocampal storage, and a transformer for semantic completion. Images are encoded, partially stored, and later reconstructed into plausible, not necessarily accurate, scenarios. The model introduces key innovations: blending multiple memory traces, incorporating emotional valence and biased reconstruction, and handling ambiguous stimuli. These features allow it to simulate SIB effects and social influences on memory, offering insights into how communication and context shape what and how we remember.

Keywords: generative episodic memory; saying is believing; shared reality; non-shared reality; episodic memory; memory trace; semantic memory.

Introduction

Episodic memory is not a static repository of past events (Bartlett, 1932) but an inherently generative process that reconstructs memories from incomplete traces completed by semantic information (Fayyaz et al., 2022). This allows filling in details, making inferences, and modifying recollections as perspectives evolve (Cheng, 2024), a view supported by cognitive psychology (Neisser, 1967). Studies on misinformation and false memories show how external suggestions and internal biases reshape memory, underscoring its adaptive, malleable nature (Loftus & Pickrell, 1995; Loftus, 2005). Constructive memory's role in social bonding is well-documented. Remembering can connect speaker and listener (Schacter, 2012; Hirst & Echterhoff, 2018), and unreliable memories support joint recollections that form shared representations (Hirst & Echterhoff, 2012), shared reality (Echterhoff, Higgins, & Levine, 2009), collective memory (Brown, Kouri, & Hirst, 2012), and social identity (Hirst & Rajaram, 2014). The saying-is-believing (SIB) paradigm (Higgins & Rholes, 1978; Echterhoff, Higgins, & Groll, 2005) demonstrates that articulating information alters the speaker's memory and attitudes. Shared reality research shows that striving for mutual understanding modifies memory to align with socially endorsed views (Echterhoff et al., 2009; Echterhoff & Higgins, 2017). Yet, computational models rarely address these social effects, focusing on individual cognition and neglecting communication-memory interactions. Generative models offer a promising path forward. We replicate audience tuning effects (Echterhoff, Higgins, Kopietz, & Groll, 2008) using a generative episodic memory model (Fayyaz et al., 2022) that forms and reconstructs traces (Cheng, 2024; Werning, 2020). It stores incomplete aspects of episodes, which are semantically completed during recall. It uses a Vector-Quantized Variational Autoencoder (VQ-VAE) (van den Oord, Vinyals, & kavukcuoglu, 2017) for compression and a BERT-based transformer (Devlin, Chang, Lee, & Toutanova, 2019) for semantic reconstruction, generating coherent images from partial traces. We extend this model by integrating multiple traces to reflect social blending, incorporating emotional valence via classifier and transformer biasing, and handling ambiguous stimuli to capture real-world complexity. With these improvements we simulate the SIB effect and provide a computational account of how verbalization and social context shape memory.

Methods and Results

The generative episodic memory model functions through a series of neural network-based steps: an input image is transformed by a VQ-VAE encoder into a latent array of convolutional feature vectors, which are then quantized into a perceptual index matrix $\mathbf{z}_q$ via codebook vector indices. A random subset of these indices is selected and stored as a memory trace using a mask that discards a fixed percentage, modeling incomplete storage due to limited cognitive resources. In parallel, the image is classified into one of ten classes, providing an evaluative judgment that biases later reconstruction. During recall, the incomplete masked trace is completed by a transformer. The transformer is trained on partially masked MNIST images and conditioned on audience judgments to reproduce the codebook vector indices, which are later decoded via the VQ-VAE decoder into a reconstructed image. This models biased generative memory and produces a remembered scenario, which can again be evaluated for valence. Although VQ-VAE and transformer models are high-level abstractions that do not capture every biological detail, they offer the right level of abstraction for our purposes. The VQ-VAE encoder mirrors feedforward visual processing, similar to the transformation of visual input into abstract object representations in the inferior temporal cortex (Yamins et al., 2014; Kuzovkin et al., 2018; Lindsay, 2021). The decoder reverses encoding, analogous to top-down feedback reconstructing cortical memory patterns (Xia, Guan, & Sheinberg, 2015; Takeda, 2019). The generated images visualize memory and do not imply literal reactivation in early visual areas. The transformer learns statistical relations among VQ-VAE features, allowing "filling in" gaps like higher cortical areas do (Tang et al., 2018; Carrillo-Reid & Yuste, 2020). Our model aligns with hippocampal indexing theory (Teyler & DiScenna, 1986), where codebook vectors connect to full features via indices. We extend this to propose that the hippocampus stores partial representations, emphasizing reconstruction over storage. Attention is modeled as random selection to simulate incomplete encoding; future versions will include input-dependent attention to better align with empirical findings. The experimental setup

we are here replicating with an extended version of the base model follows the saying-is-believing (SIB) paradigm (Higgins & Rholes, 1978; Echterhoff et al., 2005, 2008), where participants read an ambiguous text about a target person. The text was evaluatively ambiguous, containing several behavioral descriptions of the target person that could be interpreted in either a positive or a negative way. They were then asked to describe him to an audience whose attitude toward the target person was already known to them (positive/negative). In shared reality conditions, the participants had an epistemic trust in the audience's judgment and described the target so the audience could identify him. In non-shared reality conditions, the participants did not consider the audience's judgment to be reliable or credible and message production served other goals (e.g., politeness or incentives). After a distraction task, participants recalled the original text. Produced messages and recalls were blindly rated for valence on a -5 to +5 scale; the difference between audience conditions indicated the SIB effect. In place of ambiguous texts used in the experimental setup, we generate ambiguous images using a VAE trained on MNIST (LeCun, Cortes, & Burgess, 2012). It encodes images into a 2D Gaussian latent space; nearby points represent similar digits, and transitions between classes produce ambiguous samples. To quantify ambiguity, we use a 10-class Softmax classifier: images with close probabilities for two digits (e.g., 0.5 for "3", 0.4 for "8") are ambiguous. These define strong and weak labels, mapped to positive and negative judgments. Figure 1 compares simulation results (bottom) with experimental results (top) in terms of valence. Experimental findings (Echterhoff et al., 2008; Wagner, Higgins, Axmacher, & Echterhoff, 2024) show message valence depends on audience tuning and intrinsic tone. Communicators align messages with audience judgment more during message production than during recall. In non-shared reality, tuning is stronger during messaging but weaker at recall, due to conscious alignment fading over time. Control experiments (not shown) confirm that without transformer biasing, valence for ambiguous stimuli stays near zero, demonstrating that biasing is essential to reproduce observed behavior.

Conclusion and Future Work

Our study introduces a computational model for the effect of social interaction on a communicator's episodic memory. Integrating valence, biased recall, and multiple memory traces, the model operates on ambiguous stimuli, captures social effects on memory, and reproduces key experimental findings. During message production, the communicator aligns messages with audience attitudes, leading to memories shaped by communicated bias—an audience-congruent shift illustrating the saying-is-believing effect under shared reality (Echterhoff et al., 2005, 2008, 2009). While the model captures this effect already, distinguishing audience-driven changes (modeled by bias) from self-directed processes (modeled by weighting combined memory traces), as in multiple trace theory, could yield deeper insights. This work fills a critical gap by integrating communication effects into a single computational framework and lays the foundation for studying the interplay between social influence, motivation, and memory—being, to our knowledge, the first model to do so. Our modeling suggests refinements for future experiments: assessing communicator judgment before message production to quantify bias; comparing final messages to both original input and output to track false vs. accurate recall; and disentangling the influence of masking level and judgment bias in reconstruction. Since behavioral data reflect only final outcomes, it remains unclear whether memory changes stem from fading details or judgment shifts, or whether ambiguity intensifies or resolves over time. Next steps include applying the model to broader data, incorporating reaction times as indicators of cognitive accessibility (Wagner et al., 2024), and using naturalistic stimuli. While MNIST was suitable for initial validation, it lacks real-world complexity. Given the challenge of sourcing naturally ambiguous data, we plan to test the model on richer inputs. A variant has already been implemented with ImageNet-scale input (Reyhanian, Fayyaz, & Wiskott, 2024), demonstrating scalability to complex domains.

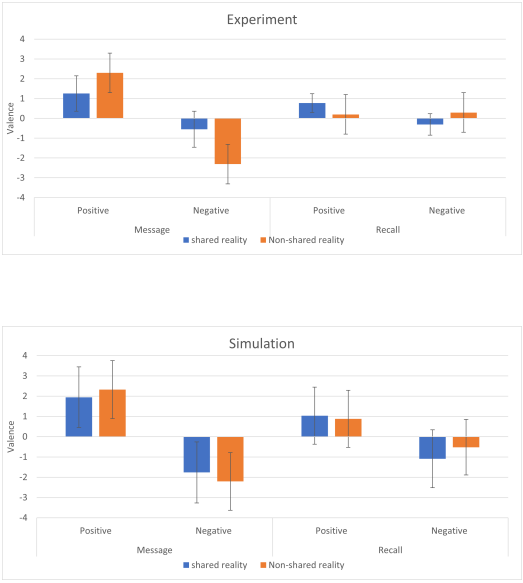

Figure 1: Simulation in comparison to experimental results.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant 397530566, FOR 2812, P5 (Wiskott/Cheng) and P9 (Echterhoff).

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Brown, A. D., Kouri, N., & Hirst, W. (2012). Memory's malleability: Its role in shaping collective memory and social identity. *Frontiers in Psychology*, *3*, 257. doi: 10.3389/fpsyg.2012.00257

Carrillo-Reid, L., & Yuste, R. (2020). Playing the piano with the cortex: Role of neuronal ensembles and pattern completion in perception and behavior. *Current opinion in neurobiology*, *64*, 89–95. doi: 10.1016/j.conb.2020.03.014

Cheng, S. (2024). Distinct mechanisms and functions of episodic memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1913), 20230411. doi: 10.1098/rstb.2023.0411

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Echterhoff, G., & Higgins, E. T. (2017). Creating shared reality in interpersonal and intergroup communication: The role of epistemic processes and their interplay. *European Review of Social Psychology*, *28*(1), 175–226. doi: 10.1080/10463283.2017.1333315

Echterhoff, G., Higgins, E. T., & Groll, S. (2005). Audience-tuning effects on memory: the role of shared reality. *Journal of Personality and Social Psychology*, *89*(3), 257–276. doi: 10.1037/0022-3514.89.3.257

Echterhoff, G., Higgins, E. T., Kopietz, R., & Groll, S. (2008). How communication goals determine when audience tuning biases memory. *Journal of Experimental Psychology: General*, *137*(1), 3–21. doi: 10.1037/0096-3445.137.1.3

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*(5), 496–521. doi: 10.1111/j.1745-6924.2009.01161.x

Fayyaz, Z., Altamimi, A., Zoellner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022). A model of semantic completion in generative episodic memory. *Neural Computation*, *34*(9), 1841–1870. doi: 10.1162/neco$_{a0}$1520

Higgins, E., & Rholes, W. S. (1978). Saying is believing: Effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology*, *14*(4). doi: https://doi.org/10.1016/0022-1031(78)90032-X

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Annual Review of Psychology*, *63*, 55–79. doi: 10.1146/annurev-psych-120710-100340

Hirst, W., & Echterhoff, G. (2018). More to episodic memory than epistemic assertion: The role of social bonds and interpersonal connection. *The Behavioral and Brain Sciences*, *41*, e17. doi: 10.1017/S0140525X17001388

Hirst, W., & Rajaram, S. (2014). Toward a social turn in memory: An introduction to a special issue on social memory. *Journal of Applied Research in Memory and Cognition*, *3*(4), 239–243. doi: 10.1016/j.jarmac.2014.10.001

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., ... Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, *1*(1), 1–12. doi: 10.1038/s42003-018-0110-y

LeCun, Y., Cortes, C., & Burgess, C. J. C. (2012). The mnist database of handwritten images. *Unpublished*.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031. doi: 10.1162/jocn$_{a0}$1544

Loftus, E. F. (2005). Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *12*(4), 361–366. doi: 10.1101/lm.94705

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725. (Original work published December 1, 1995) doi: 10.3928/0048-5713-19951201-07

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Reyhanian, S., Fayyaz, Z., & Wiskott, L. (2024, September). Analysis of a generative model of episodic memory based on hierarchical vq-vae and transformer. In *Proceedings of the 33rd international conference on artificial neural networks (icann 2024), lugano, switzerland.* Springer Nature Switzerland. doi: 10.1007/978-3-031-72341-4$_6$

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *The American Psychologist*, *67*(8), 603–613. doi: 10.1037/a0029869

Takeda, M. (2019). Brain mechanisms of visual long-term memory retrieval in primates. *Neuroscience research*, *142*, 7–15. doi: 10.1016/j.neures.2018.06.005

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840. doi: 10.1073/pnas.1719397115

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral neuroscience*, *100*(2), 147.

van den Oord, A., Vinyals, O., & kavukcuoglu, k. (2017). Neural discrete representation learning. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Wagner, U., Higgins, E. T., Axmacher, N., & Echterhoff, G. (2024, Jun). Biased memory retrieval in the service of shared reality with an audience: The role of cognitive accessibility. *Journal of Experimental Psychology: General*, *153*(6), 1605–1627. doi: 10.1037/xge0001580

Werning, M. (2020). Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology*, *11*(2), 301–333. doi: 10.1007/s13164-020-00471-z

Xia, R., Guan, S., & Sheinberg, D. L. (2015). A multilayered story of memory retrieval. *Neuron*, *86*(3), 610–612. doi: 10.1016/j.neuron.2015.04.017

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111