

# AUGUSTUS: An LLM-Driven Multimodal Agent System with Contextualized User Memory

Jitesh Jain<sup>1\*</sup> Shubham Maheshwari<sup>2\*†</sup> Ning Yu<sup>3</sup> Wen-mei Hwu<sup>4</sup> Humphrey Shi<sup>1</sup>

<sup>1</sup>SHI Labs @ Georgia Tech <sup>2</sup>Adobe <sup>3</sup>Netflix Eyeline Studios <sup>4</sup>UIUC

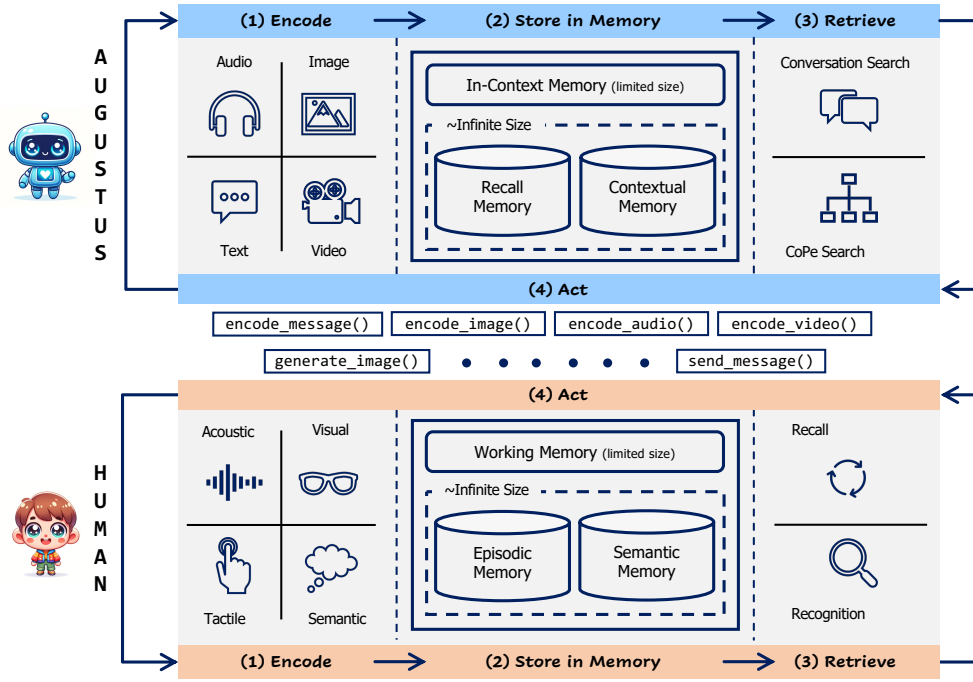


Figure 1: **Alignment with Human Cognition System.** AUGUSTUS first encodes information for storage into long-term (external) memory, using the limited context length in LLMs as working memory and retrieving information upon the need to act.

## Abstract

Riding on the success of LLMs with retrieval augmented generation (RAG), there has been a growing interest in augmenting agent systems with external memory databases. However, the existing systems focus on storing text information in their memory, ignoring the importance of multimodal signals. Motivated by the multimodal nature of human memory, we present **AUGUSTUS**, a multimodal agent system aligned with the ideas of human memory in cognitive science. Technically, our system consists of 4 stages connected in a loop: (i) *encode*: understanding the inputs; (ii) *store in memory*: saving important information; (iii) *retrieve*: searching for relevant context from memory; and (iv) *act*: perform the task. Unlike existing systems that use sequential key-value databases, we propose conceptualizing information into semantic tags and associating the tags with their context to store them in a graph-structured multimodal contextual memory for efficient concept-driven retrieval. Our system outperforms the traditional multimodal RAG approach while

\*Equal Contribution. †Work done during SM's time at Georgia Tech.

being 3.5 times faster for ImageNet classification and outperforming MemGPT on the MSC benchmark.

## 1 Introduction

Processing information is a crucial part of a human’s existence [1]. At the core of processing lies the memory system [2] in humans, persistently storing, retrieving, and modifying multimodal information across dynamic working and long-term memory. The human ability to react to every situation based on the context is guided by memory to a large extent [3, 4]. Recently, LLMs have emerged as agent systems [5–7] that think [8, 9]. Although impressive, we find that the role of memory in these agent systems is relatively less studied, even less so for multimodal memory. We observe that most existing systems either maintain a vector database knowledge base, performing retrieval augmented generation (RAG) when necessary [10, 11] or do not maintain or access any form of external memory, therefore, only capable of planning the execution of tools defined in their system prompt [5, 12]. More recently, a few works [7, 13, 14] proposed maintaining a dynamic long-term memory, which is updated as the agent interacts with other agents or the user. However, the mentioned systems (and their long-term memory) only support the text modality during their interactions with a user. Intuitively, an agent system should be able to retain all context from user conversations, including any relevant image, audio, or video data for future conversations. Therefore, developing an agent system that can process, store, and retrieve multimodal information is imperative [15].

Working towards such a system, we present **AUGUSTUS**, a multimodal agent storing user context in memory, seeking answers to the following questions:

(i) **Is it possible to establish an analogy between existing agent systems and the human cognition system?** Based on past research in cognitive neuroscience, working memory is considered to have limited capacity [16]. In contrast, long-term memory is believed to have unlimited capacity [17]. We recognize the similarity between the working memory in humans and the in-context memory in large language models (LLMs), with both having a limited size and being the primary source for planning/reasoning for making a decision [18]. We observe another analogy between the cue-driven retrieval [19] in humans from long-term memory and RAG [20] using external memory in LLMs. Lastly, we focus on two main types of long-term memory [21]: *episodic memory*, responsible for storing temporally dated information, and *semantic memory*, responsible for storing concepts and relation among the held concepts. Recently, Packer *et al.* [13] proposed using a *recall memory* to store all user (text) conversations in chronological order. We realize the similarities between episodic and recall memory while developing AUGUSTUS. Moreover, we equip AUGUSTUS with a contextual memory to store conceptualized information, similar to semantic memory. We present more details on the analogy in Appendix A.

(ii) **Building upon the observed analogies, is it possible to develop a general multimodal agent system good at remembering context about the user?** Recently, there has been a growing interest in developing any-to-any systems [22–25] due to their modular design. We believe a general system should be able to understand and generate multiple modalities. Therefore, we also leverage existing foundation models [26–29] as tools to equip AUGUSTUS to understand and generate data for multiple modalities. At the center of our system lies an LLM, the planner responsible for creating function calls to execute the corresponding tool. We also provide our system with the appropriate tools to store and retrieve information in/from two types of external (long-term) memory databases: recall [13] and contextual memory. We attempt to imitate the humans’ ability to conceptualize information [21] for storage in semantic memory by equipping our agent with the ability to abstract information into semantic tags (concepts) to store in the contextual memory. The tags in the contextual memory are connected to context nodes that store information about the user from the corresponding conversation snapshot [30], including any media. Thus, the contextual memory stores the multimodal context about the user. Additionally, we introduce the Contextual-Personalized (CoPe) search mechanism to perform concept-driven retrieval of information from the contextual memory to provide personalized responses (Fig. 5) to the user.

To summarize, our contributions are three-fold:

- We establish an analogy between human cognition and agent systems, introducing AUGUSTUS with a contextual memory aligned with the ideas of a multimodal memory system in humans from cognitive neuroscience.

Table 1: **AUGUSTUS’ Tool Collection.** We equip our system with various tools for smooth multimodal operation. Our experiments use tools in **bold** as default choices.

| Stage           | Tool                                 | Modality                  | Model Access   |
|-----------------|--------------------------------------|---------------------------|----------------|
| Encode          | <b>Video-LLaVA</b> [34]              | image; video              | Open-Source    |
|                 | <b>WhisperX</b> [41]                 | audio ( <i>speech</i> )   | Open-Source    |
|                 | <b>Whisper-Captioner</b> [27]        | audio                     | Open-Source    |
|                 | <b>dolphin-2.6-mixtral-8x7b</b> [40] | text                      | Open-Source    |
|                 | GPT-4 [10]                           | text                      | Close-Source   |
| Store in Memory | <b>Recall Memory (sqliteDB)</b>      | text                      | Not Applicable |
|                 | <b>Contextual Memory (arangoDB)</b>  | image; video; audio; text | Not Applicable |
| Generate        | <b>SD-2.1</b> [28]                   | image                     | Open-Source    |
|                 | <b>MGIE</b> [36]                     | image ( <i>edit</i> )     | Open-Source    |
|                 | <b>LaVIT</b> [42]                    | image ( <i>edit</i> )     | Open-Source    |
|                 | <b>MAGNeT</b> [43]                   | audio                     | Open-Source    |
|                 | <b>I2VGen-XL</b> [29]                | video                     | Open-Source    |
|                 | <b>dolphin-2.6-mixtral-8x7b</b> [40] | text                      | Open-Source    |
|                 | GPT-4 [10]                           | text                      | Close-Source   |

- We equip our agent system with the ability to conceptualize information for storage in contextual memory for a more efficient memory system. Moreover, we introduce a two-stage CoPe search for effective concept-driven retrieval.
- We provide empirical evidence on the effectiveness of our system, outperforming the traditional multimodal RAG approach while being 3.5 times faster. Moreover, AUGUSTUS outperforms MemGPT [13] on the MSC benchmark.

## 2 Related Works

### 2.1 LLM-Driven Agent Systems

With the rise of LLMs for various language and multimodal tasks [26, 31], the community has leveraged LLMs as planners [32, 33] in agent systems, having three modules: *tool collection*, *planning module*, and *memory*.

The *tool collection* includes multimodal foundation models for understanding the user query [27, 34, 35] and generating the response [28, 29, 36].

The *planning module* is an LLM that generates a sequence of execution steps using its reasoning ability to fulfill the user query. The format and instructions for the step generation are passed as few-shot or zero-shot in-context prompt to the LLM to achieve successful step generation, which can either be a natural language [8, 37] or programs [5], or both [13, 38].

The community has incorporated long-term external memory databases into agent systems [14, 39] to overcome the LLMs’ limited context length.

For AUGUSTUS, we also use off-the-shelf foundation models during the *encode* and *act* stages. For planning, inspired by the success and elegance of planning through function calling in MemGPT [13], we also follow the same strategy as a combination of natural language and programs. We use a fine-tuned version of Mixtral-8x7b [26, 40] as our planning module for improved function calling abilities. AUGUSTUS can directly access the contents in the external memory, making our system’s memory management completely autonomous [13].

### 2.2 Agent Systems with Memory

Inspired by the success of retrieval-augmented-generation (RAG) [20, 44], existing agent systems have also incorporated an external memory in vector databases to overcome the limitations of limited context in LLMs. Park *et al.* [7] maintained agent-specific memory to store the experiences of each persona as different agents interacted with each other in a simulation. MemGPT [13] proposed holding two types of external databases: storing raw conversation history in its recall memory and essential information about the user in archival memory. OS-Copilot [14] proposed a more cognitively aligned approach, storing the user and the agent’s action history information

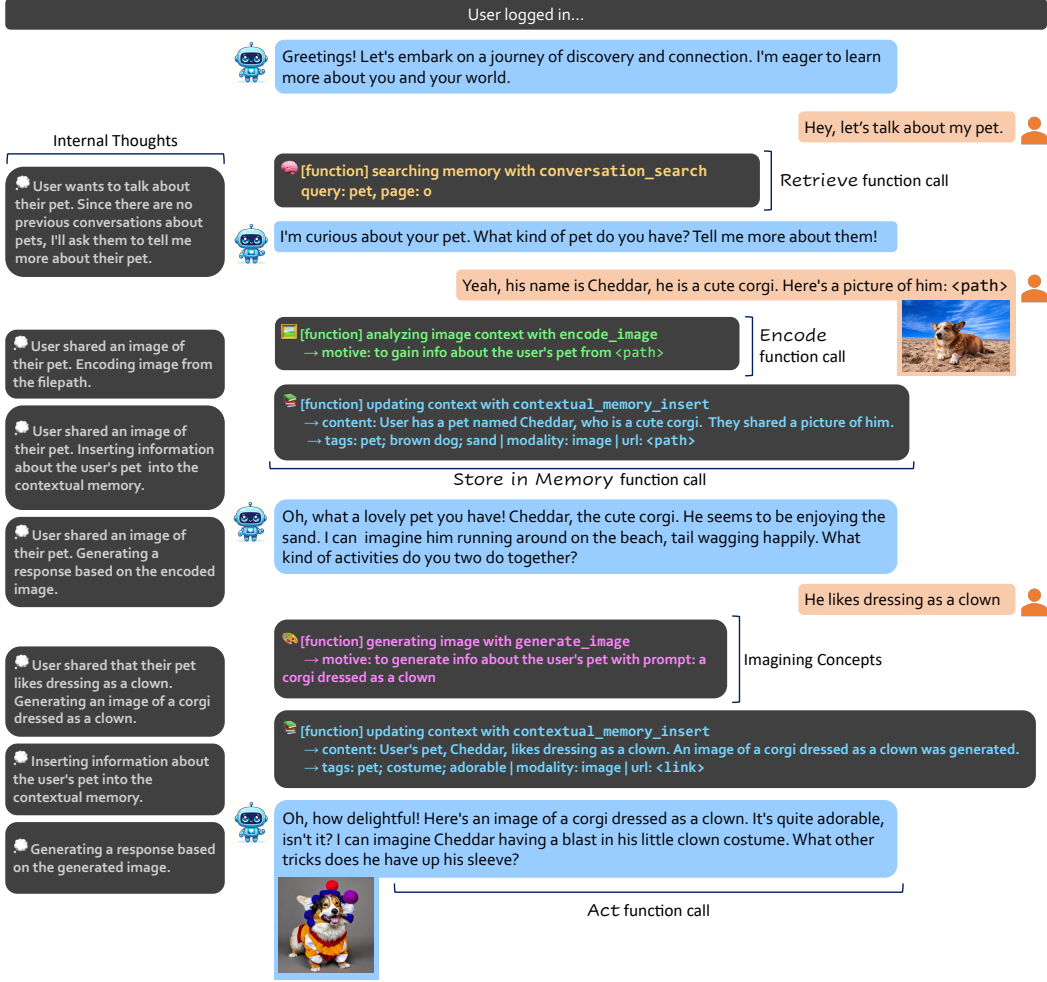


Figure 2: **Snapshot of a conversation between a user and AUGUSTUS about the user’s pet.** Our system calls different functions corresponding to the four stages of operation to engage with the user while thinking internally [8] for careful planning.

in a declarative memory database while storing information about tools in a procedural memory database to support learning new tools for acting as an operating system. Although impressive, these systems only support storing text information in their external memory databases. More recently, JARVIS-1 [39] proposed storing both the plan and images in memory to improve performance at solving tasks in the Minecraft game environment. However, JARVIS-1 maintained a non-hierarchical memory while only supporting image and text modalities, not aligning with the human memory system (Appendix A), leading to inefficient retrieval with increased memory size. In contrast to the existing agent systems, our AUGUSTUS maintains the contextual memory as a hierarchical multimodal database, storing semantic concepts at the top level with the associated context stored under the corresponding concepts, reducing the search space.

### 3 System Description

Inspired by the ability of humans to leverage context during social interactions based on experience and knowledge from memory [45], we present AUGUSTUS, an agent system aligned with the ideas of the human memory system in cognitive science (Appendix A). As shown in Fig. 1, we design our system with four stages connected in a loop: (i) encode, (ii) store in memory, (iii) retrieve, and (iv) act. Following previous works [7, 13, 38], we also equip our agent system with the ability to think [8, 46] for more accurate tool selection and function calling. We demonstrate a snapshot of AUGUSTUS’ conversation with a user in Fig. 2 during which our agent tries to understand and store



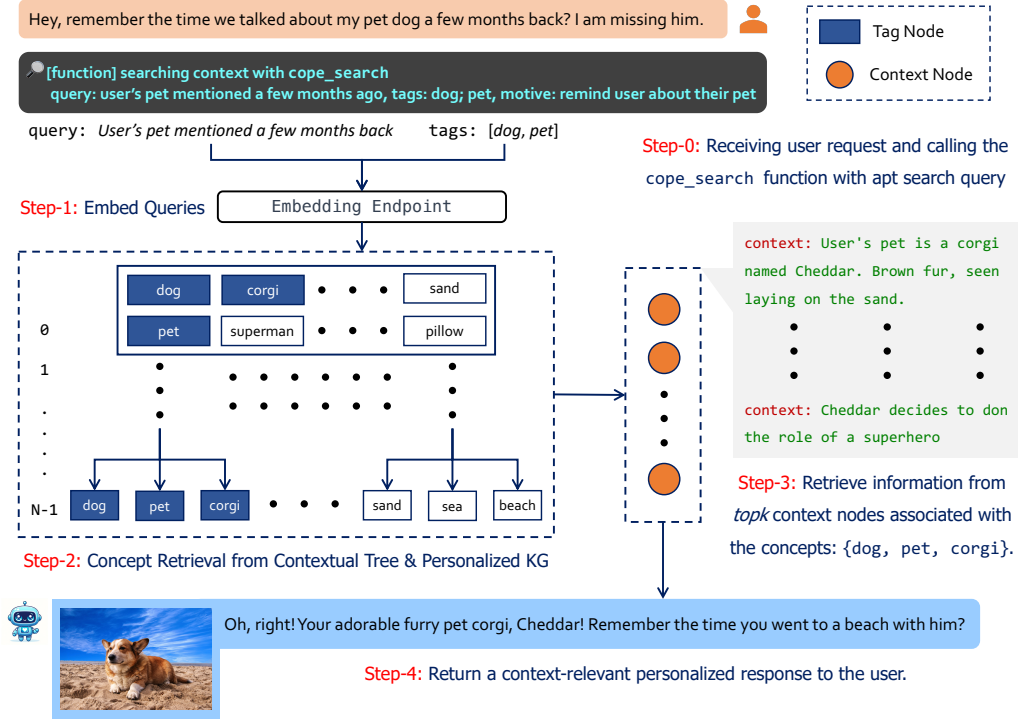


Figure 4: **Retrieval with CoPe Search.** Given a query, AUGUSTUS retrieves relevant concept (*pet*; *dog*; *corgi*). followed by concept-driven context retrieval to send a personalized response to the user (name and image of the *pet* to help user reminisce).

**Retrieve.** Our system’s retrieval is completely autonomous, with the agent deciding when to search for information. AUGUSTUS retrieves information from the recall memory with a paginated search over the storage for the query [13]. To retrieve information from the contextual memory, we formulate a two-stage search approach, first searching for relevant tags (semantic abstraction of context) and then searching over the context nodes associated with the retrieved tags to return personalized information relevant to the present conversation (Fig. 5). We term our algorithm **Contextual-Personalized (CoPe)** search.

**Act.** We equip AUGUSTUS with a range of actions (Tab. 1), including sending a message, understanding and generating an image, audio, and video, and the ability to edit images. Note that different actions can be chained together, with sending a message to the user always being the last action in the chain.

### 3.1 Storing Concepts in Contextual Memory

As shown in Fig. 3, the information inside the contextual memory is organized in a graph structure with associations among tags based on shared context. We store the concept name inside a *tag node* and the multimodal user context inside a *context node*. If two tag nodes share a context node, they must have an edge; if two tag nodes have an edge, they must share at least one context node.

During a conversation, AUGUSTUS initiates the information insertion process in the contextual memory with a `contextual_memory_insert` function call. The first step is conceptualizing the encoded information into semantic tags. Next, depending on the modality of the information, our system pushes non-text information to cloud storage and stores the corresponding URL, a list of tags, encoded information, modality type, and timestamp in a context node object, as shown at the top in Fig. 3. Note that for text modality, the URL field is null. To ensure the high quality of generated tags, we leverage LLMs’ in-context-learning (ICL) [47] ability (LLMs) and pass demonstrations for improved conceptualization. We share the demonstrations in the appendix.

Table 2: **Concept-Retrieval for the ImageNet-1k classification task.** Our system shows comparable performance to the specialized classification frameworks, showing the effectiveness of our contextual memory and CoPe search.

| Method                                | Trained on ImageNet | Retrieval | Top-1 ( $\uparrow$ ) |
|---------------------------------------|---------------------|-----------|----------------------|
| MobileNetV2 [50]                      | ✓                   | ✗         | 72.0                 |
| Description based Classification [51] | ✗                   | ✗         | 76.2                 |
| Vit-L/16 [52]                         | ✓                   | ✗         | 76.5                 |
| ResNet-50 [53]                        | ✓                   | ✗         | 77.6                 |
| ImageBind [54]                        | ✓                   | ✗         | 77.7                 |
| ViT-B/16 [52]                         | ✓                   | ✗         | 77.9                 |
| Perceiver [55]                        | ✓                   | ✗         | <b>78.0</b>          |
| <b>AUGUSTUS (ours)</b>                | ✗                   | ✓         | 75.0                 |
| <b>AUGUSTUS (ours) w/o clustering</b> | ✗                   | ✓         | <b>78.0</b>          |

### 3.2 Retrieve from Contextual Memory: CoPe search

Given a `cope_search` function call, our system first searches over the concepts and then retrieves the concept-relevant context to respond to the user, as shown in Fig. 4. Naturally, the search becomes more compute-intensive as the number of tags grows. Therefore, we construct a *contextual tree* of tags to perform an efficient search over many concepts, with each node representing the average of context embeddings associated with the corresponding tag. Owing to the high dimensionality of embeddings, we perform UMAP [48] over tag features and iteratively cluster tags with HDBSCAN [49] until there’s a single cluster as the root node. As mentioned in Sec. 3.1, in our contextual memory, we associate concepts based on shared context nodes, much like a *personalized (knowledge) graph* structurally dependent on user context information from their interactions with AUGUSTUS. We leverage these associations to further boost our concept retrieval performance by learning concept representations about the user context to predict a set of personalized tags. We collectively use the two sets of tags for the context retrieval step, returning the context nodes based on their cosine similarity with the query embeddings. We only compare the similarity values for the concept-relevant context nodes, reducing the search space, unlike searching through all the nodes in a sequential vector database [13]. We use the same embedding model to compute the embeddings for the tags, query, and multimodal context nodes. We provide our CoPe search logic in Algo. 1.

### 3.3 Function Calling

The operation of AUGUSTUS is driven by an LLM through a function calling mechanism [5, 13]. We include the description of functions responsible for the four stages in our agent as a part of our system prompt instructions to the LLM. During a conversation, the LLM generates the values for the parameters to be passed into the corresponding function for smooth operation. For example, the `send_message` function expects a *string* (message) data type; therefore, the LLM shall pass a *string* to the corresponding function call. We list all the supported functions in Appendix E.

## 4 Experiments

**Implementation Details.** Unlike existing agent systems [5, 13, 39] that usually leverage OpenAI API access-based models, we instead prioritize using open-source models as our tools for better reproducibility and low financial costs. Our system also supports API-based models, as shown in Tab. 1. As our LLM, we use the `dolphin-2.6-mixtral-8x7b` [40] model. While retrieving information from contextual memory, we add the support for ImageBind [54] and LanguageBind [56] (default) models to compute the embeddings for similarity calculation between the search query and information stored in context nodes. Note that for evaluation purposes, we use ImageBind [54] as our embedding model.

**Evaluation Datasets.** We evaluate the effectiveness of our CoPe search algorithm for multimodal retrieval on the ImageNet [57] ILSVRC-2012 classification task. Moreover, we also quantify our system’s performance as a conversational agent on the Multi-Session Chat (MSC) benchmark [58] to measure our system’s ability to answer questions based on prior user conversations.

Table 3: **Ablations on various retrieval techniques.** Our CoPe search demonstrates the best performance-to-efficiency trade-off (in **bold**) with comparable accuracy to standard multimodal RAG while being nearly 3.5 times faster for same memory size.

| Method                    | tag clustering? | query   | Memory Size (approx.) | Top-1 (↑)   | Top-5 (↑)   | time (ms) (↓) |
|---------------------------|-----------------|---------|-----------------------|-------------|-------------|---------------|
| Multimodal RAG            | N/A             | image   | 10k nodes             | 62.3        | 86.8        | 113.3         |
| Multimodal RAG            | N/A             | image   | 100k nodes            | 71.7        | 92.1        | 137.1         |
| Multimodal RAG            | N/A             | image   | 1200k nodes           | 78.2        | 95.2        | 403.8         |
| CoPe Search (ours)        | ✓               | caption | 1200k nodes           | 27.0        | 45.4        | 108.2         |
| CoPe Search (ours)        | ✗               | caption | 1200k nodes           | 46.8        | 74.4        | 110.1         |
| <b>CoPe Search (ours)</b> | ✓               | image   | 1200k nodes           | <b>75.0</b> | <b>91.1</b> | <b>99.1</b>   |
| <b>CoPe Search (ours)</b> | ✗               | image   | 1200k nodes           | <b>78.0</b> | <b>94.8</b> | <b>126.1</b>  |

#### 4.1 Concept Retrieval on ImageNet

To show the effectiveness of our CoPe search algorithm at retrieving relevant semantic tags from the contextual memory, we perform multimodal retrieval for the ImageNet-1k [57] classification task. Specifically, we first populate our contextual memory with ImageNet class names as *tag nodes* and the corresponding images from the training set attached as *context nodes*, producing 997 tag nodes and about 1.2M context nodes as the search space in our contextual memory. Note that we only have 997 tag nodes (and not 1000) due to three duplicate class names in ImageNet-1k: *crane*, *cardigan*, and *maillot*. We experiment under two settings for CoPe search: default search with tag clustering (number of nodes being 1, 6, and 997 at different levels) and CoPe search without clustering, essentially searching over all 997 tags. We use images from the validation set (containing 50k samples) as our search query.

As shown in Tab. 2, our system shows impressive retrieval performance with top-1 accuracy. Moreover, when we turn off tag node clustering during the retrieval, our system outperforms established classification frameworks like ViT [52], demonstrating the effectiveness of our CoPe search mechanism.

We also analyze the effect of search query and database type on the accuracy and retrieval time. As shown in Tab. 3, using image as the search query shows the best performance, demonstrating the importance of supporting multimodal retrieval from the contextual memory. We obtain captions using Video-LLaVA [34] for the search query. Moreover, we ablate with standard multimodal RAG [59] over a vector database storing images and class names as key-value pairs. Our CoPe search performs comparably to multimodal RAG and is nearly 3.5 times faster, reaffirming our claim that our concept-driven search is more efficient than searching over key-value pair-based vector databases for large memory sizes.

#### 4.2 Conversation Consistency on MSC

We evaluate AUGUSTUS’ performance at remembering context as a conversational agent on the MSC [58] dataset. Before evaluation, we populate our contextual memory with the context from the conversation history from the multiple sessions from the validation set containing 500 multi-session conversation samples. We use the QnA pairs proposed in [13] for the deep memory retrieval task to assess the benefit of contextual memory. As shown in Tab. 4, AUGUSTUS outperforms other methods on the ROUGE-L (R) [60] metric, demonstrating the effectiveness of our system as a conversational agent. The LLM-only baselines see a summary of prior conversations as context prompts during inference [13].

## 5 Conclusion

In this work, we presented AUGUSTUS, a multimodal agent system aligned with the principles of human memory with four processing stages connected in a loop. Inspired by the semantic memory in humans, we propose a hierarchical contextual memory for storing multimodal information as semantic concepts associated with shared context from user conversations. We devise the Contextual-Personalized (CoPe) search algorithm for efficient concept-driven retrieval from contextual memory. Our experiments demonstrate the effectiveness of AUGUSTUS over existing retrieval techniques and conversational agents while being more efficient than systems with sequential databases as the size of

Table 4: **Performance as a conversational agent.** AUGUSTUS outperforms other baselines on the deep memory retrieval task on the MSC [58] benchmark, showing the effectiveness of our system. \* denotes results taken directly from Packer et al. [13].

| Method                                    | ROUGE-L (R) (↑) |
|-------------------------------------------|-----------------|
| <i>LLM-only</i>                           |                 |
| GPT-3.5* Turbo                            | 0.394           |
| GPT-4*                                    | 0.296           |
| GPT-4-Turbo*                              | 0.359           |
| dolphin-2.6-mixtral-8x7b                  | 0.280           |
| <i>Agent Systems with Memory</i>          |                 |
| MemGPT [13] with dolphin-2.6-mixtral-8x7b | 0.405           |
| <b>AUGUSTUS (ours)</b>                    | <b>0.456</b>    |

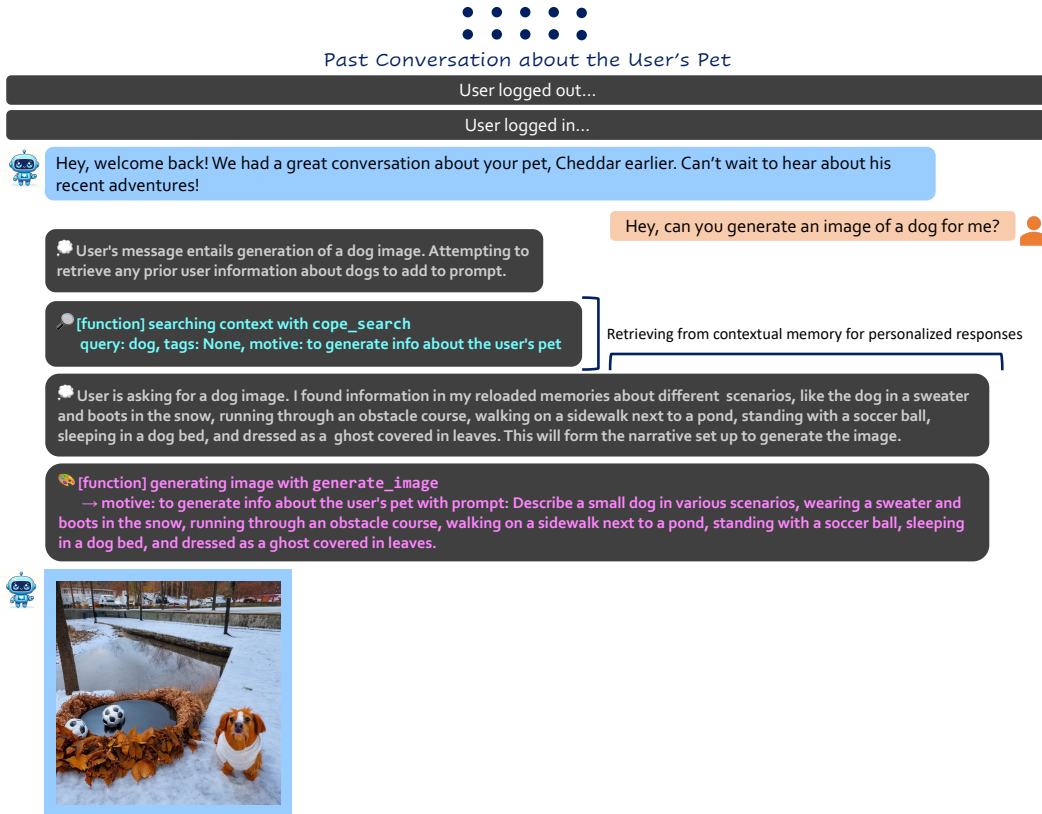


Figure 5: **Personalized response to the user.** AUGUSTUS retrieves relevant information from the contextual memory by calling the `cope_search` function to generate an image aligned with the information about dogs from prior user conversations. Note that the image quality depends on our generator, SD-2.1 [28] in this case.

memory increases. We hope AUGUSTUS inspires the research community to develop future agent systems with cognition-aligned hierarchical multimodal memory.

**Limitations.** Despite our progress towards developing a more human cognition-aligned agent system, certain limitations are left to be addressed for future work. Firstly, with an LLM as our system’s planner, we sometimes observed inconsistencies between our agent’s thoughts and actions, a well-known issue with chain-of-thought prompting [61]. Although rare, solving these inconsistencies is crucial in developing more dependable and cognition-aligned agent systems. Secondly, equipping AUGUSTUS with the ability to learn new tools and generate novel function calls [14, 39] given a set

of tools is a promising future direction. Lastly, it is crucial to develop benchmark datasets to evaluate multimodal conversational agents on their context-retaining ability effectively in the future.

**Acknowledgements.** We extend our heartfelt gratitude to Pratyusha Maiti and Jeremy Collins for their insightful discussions. This work was in part supported by NSF CAREER Award #2239840, and the National AI Institute for Exceptional Education (Award #2229873) by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education. Lastly, we thank the ML Center @Georgia Tech for supporting this work.

## References

- [1] Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience: The Biology of the Mind*. W.W. Norton & Company, 5 edition, 2018. 2
- [2] Derek Bok Center for Teaching and Learning. How memory works, 2023. Accessed: 2024-02-26. 2, 15
- [3] Daniel L. Schacter and Endel Tulving. The neuropsychology of memory. *Neuropsychology*, 11(1):3, 1997. 2
- [4] Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. *Principles of Neural Science*. McGraw-Hill, 4th edition, 2000. 2
- [5] Surís Dídac, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICLR*, 2023. 2, 3, 5, 7, 16
- [6] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv*, 2023.
- [7] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023. 2, 3, 4, 5
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 2023. 2, 3, 4, 5, 15
- [9] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. *arXiv*, 2023. 2
- [10] OpenAI. Gpt-4 technical report, 2023. 2, 3
- [11] Google Research. Gemini: A family of highly capable multimodal models. *arXiv*, 2023. 2
- [12] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023. 2, 15
- [13] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards llms as operating systems. *arXiv*, 2023. 2, 3, 4, 5, 6, 7, 8, 9, 15, 16, 17
- [14] Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. Os-copilot: Towards generalist computer agents with self-improvement. *arXiv*, 2024. 2, 3, 9
- [15] A. D. Baddeley, M. W. Eysenck, and M. C. Anderson. *Memory*. Psychology Press, New York, NY, 7th edition, 2014. General overview of memory systems, including discussions on multimodal aspects of memory. 2
- [16] Baiwei Liu, Xinyu Li, Jan Theeuwes, and Benchi Wang. Long-term memory retrieval bypasses working memory. *NeuroImage*, 261:119513, 2022. ISSN 1053-8119. 2

- [17] Nelson Cowan. What are the differences between long-term, short-term, and working memory? In *Progress in Brain Research*, volume 169, pages 323–338. Elsevier, 2008. 2
- [18] N. Cowan. Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2):197–223, 2014. 2
- [19] E. Tulving and D. M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973. doi: 10.1037/h0020071. 2, 15
- [20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*, 2021. 2, 3
- [21] Endel Tulving. Episodic and semantic memory. *Organization of Memory*, 1:381–403, 1972. 2, 15
- [22] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 2
- [23] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, 2023. 15
- [24] Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, Xiaohua (Michael) Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning. *arXiv*, 2024.
- [25] Xinyu Wang, Bohan Zhuang, and Qi Wu. Modaverse: Efficiently transforming modalities with llms. *arXiv*, 2024. 2
- [26] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mixtral of experts. *arXiv*, 2024. 2, 3
- [27] Marek Kadl  k, Adam H  jek, J  rgen Kieslich, and Rados  aw Winiecki. A whisper transformer for audio captioning trained with synthetic captions and transfer learning. *arXiv*, 2023. 3
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj  rn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 9, 16
- [29] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv*, 2023. 2, 3, 16
- [30] John R. Anderson. Act: A simple theory of complex cognition. *American Psychologist*, 51:355, 1996. 2
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3
- [32] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023. 3
- [33] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv*, 2023. 3
- [34] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv*, 2023. 3, 8

- [35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023. 3
- [36] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *ICLR*, 2024. 3
- [37] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models, 2023. 3
- [38] DiFei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv*, 2023. 3, 4
- [39] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv*, 2023. 3, 4, 7, 9, 16
- [40] TheBloke. dolphin-2.6-mixtral-8x7b-awq. <https://huggingface.co/TheBloke/dolphin-2.6-mixtral-8x7b-AWQ>, 2024. 3, 7, 17
- [41] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH*, 2023. 3
- [42] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv*, 2023. 3
- [43] Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. *arXiv*, 2024. 3, 16
- [44] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*, 2024. 3
- [45] Harold Garfinkel. *Studies in Ethnomethodology*. Prentice-Hall, 1967. 4
- [46] Michael Rescorla. The language of thought hypothesis, 2023. The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.). 4, 5, 15
- [47] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv*, 2020. 6, 21
- [48] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, Sep 2018. doi: 10.21105/joss.00861. 7

- [49] Ricardo J. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, page 160–172, 2013. doi: 10.1007/978-3-642-37456-2\_14. 7
- [50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7
- [51] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 7
- [52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7, 8
- [53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7
- [54] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 7
- [55] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021. 7
- [56] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024. 7
- [57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 7, 8
- [58] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*, 2022. 7, 8, 9
- [59] Wenhua Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *EMNLP*, 2022. 8
- [60] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004. 8
- [61] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv*, 2023. 9
- [62] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *The psychology of learning and motivation*, volume 2, pages 89–195. Academic Press, 1968. 15
- [63] Alan Baddeley. Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003. 15
- [64] Alan D Baddeley and Graham Hitch. Working memory. In *The psychology of learning and motivation: Advances in research and theory*, volume 8, pages 47–89. Academic Press, New York, 1974. 15
- [65] L. R. Squire. *Memory and Brain*. Oxford University Press, 1987. 15
- [66] Safoora Yousefi, Leo Betthausen, Hosein Hasanbeig, Raphaël Millière, and Ida Momennejad. Decoding in-context learning: Neuroscience-inspired analysis of representations in large language models. *arXiv*, 2024. 15

- [67] Howard Eichenbaum and Neal J. Cohen. *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford University Press, 2001. 15
- [68] Bence Nanay. Mental imagery. <https://plato.stanford.edu/entries/mental-imagery/>, 2021. Accessed: 2024-03-02. 16

# Appendix

In this appendix, we first share details about the alignment of AUGUSTUS’ system design with the human cognition system in Appendix A. Next, we share the system prompt for AUGUSTUS in Appendix C. Secondly, we present information about the demonstration examples for accurate conceptualization of information into semantic tags in Appendix D. Lastly, we provide details about all supported function calls in our multimodal agent systems in Appendix E.

## A Alignment with Human Cognition

We observe that the cognitive science community broadly divides the human memory system into three stages [2, 21, 62]: (i) *encode* sensory inputs; (ii) *store* in memory; and (iii) *retrieve* information. We also formulate AUGUSTUS along similar lines, aligning our system (memory) design with human cognition. Moreover, we term the last stage attributed to the tool execution in AUGUSTUS and action/reaction in humans as *act*. All four stages can be considered connected in a closed loop form, with the *act* stage providing new information to the *encode* stage, as shown in Fig. 1.

### A.1 Stage-1: Encode

**Humans.** Encoding information perceived through the sensory organs is the initial step [2] in the human memory system. It involves converting the visual, acoustic, tactile, and semantic signals into a representation storable in memory.

**AUGUSTUS** Our system also encodes the user inputs as the first step. AUGUSTUS can handle inputs across multiple modalities: image and video (visual), audio (acoustic), and text (semantic). We do not support touch (tactile) modality; however, we believe supporting touch modality is straightforward, which we leave for future work. We use modality-specific encoders to encode inputs, following the existing any-to-any works [23]. As an LLM lies at the core of our system, we convert all non-text inputs into text caption, following the language of thought hypothesis [8, 13, 46], for the LLM to reason and make decisions in our system. We provide more details about our encoders in Sec. 3. Note that, in our contextual memory, we store both the text caption and the corresponding non-text input media to prevent any loss in multimodal context.

### A.2 Stage-2: Store in Memory

**Humans.** During the second stage, the encoded information is forwarded to the memory for storage and manipulation. From the perspective of our system design, we focus on two main memory types in humans: (i) *working memory* has a limited capacity and is responsible for information manipulation and reasoning, along with other tasks that require planning and decision making [63, 64]; and (ii) *long-term memory* is believed to have unlimited storage capacity, storing information indefinitely which needs to be retrieved for use [2, 19]. We take inspiration from explicit long-term memory [65] comprised of semantic and episodic memory, with the former storing information about concepts, mainly linguistic, and the latter storing chronological information [21].

**AUGUSTUS.** With an LLM as our system’s orchestrator, we recognize the similarities between the in-context memory in LLMs and the working memory in humans. Like working memory, in-context memory also has a limited size (context length) in LLMs. Moreover, the prompt passed as the context to the LLM can be used to affect the response behavior of LLMs, imitating the information manipulation and reasoning capacity in humans [12, 66]. We implement the long-term memory in our system with two external databases: (i) *recall memory* stores the complete conversation log in raw text form, following [13], and (ii) *contextual memory* is a hierarchical multimodal database storing information in the form of semantic concepts associated [67] with each other through shared context from the corresponding conversation snapshot.

### A.3 Stage-3: Retrieve

**Humans.** In cognitive science, the encoding specificity principle [19] proposed that memory stores information with its context; therefore, retrieval is most effective when the cues present at the time

of retrieval match those present at the time of storage. Motivated by this principle, we also store concepts associated with their context in our contextual memory for effective context-driven retrieval.

**AUGUSTUS.** While retrieving information from the contextual memory, given a user query, we perform a search over the stored semantic concepts represented with their associated context in the embedding space. Next, we retrieve the information associated with the corresponding concepts for the final response.

---

**Algorithm 1: Contextual-Personalized (CoPe) Search**

---

```

1 Function cope_search(query, topk,  $\mathcal{M}$ ):
2   get contextual tree from memory  $CT \leftarrow \text{get\_CT}(\mathcal{M})$ 
3   get personalized graph from memory  $PG \leftarrow \text{get\_PG}(\mathcal{M})$ 
4   init values for tree traversal  $root \leftarrow CT.root$ , retrieved tags  $T \leftarrow []$ 
5   while True do
6     if  $root.is\_leaf\_parent()$  then
7       // retrieve the topk children concepts & break loop
8        $T.extend(CT.get\_children(root, query, topk))$ 
9       break
10    else
11      // traverse down to the top child node
12       $root \leftarrow CT.get\_top\_child(root, query)$ 
13    end
14  end
15  // obtain personalized concepts & context
16   $T.extend(PG.get\_tags(T))$ 
17   $CN \leftarrow [tag.get\_context() \text{ for } tag \text{ in } T]$ 
18  return  $T, CN$ 

```

---

#### A.4 Stage-4: Act

The final stage in our system is providing the user with an appropriate response. We equip AUGUSTUS with various generation tools [28, 29, 43] to handle different response formats for a user query. We relate the final stage of our system to how humans react to different situations with speech, text, drawing, etc.

## B Qualitative Analysis

Due to the unavailability of any multimodal dataset to evaluate user personalization performance, we qualitatively analyze our system’s ability to retrieve information from the contextual memory to provide a more personalized response. As shown in Fig. 5, when prompted to “generate an image of a dog” by the user, AUGUSTUS searches for context relevant to the “dog” in its contextual memory, retrieving dog-related information from previous conversations, consequently, generating an image aligned with context about “dogs” in memory.

As shown in Fig. 2, AUGUSTUS stores user information like “User’s pet, Cheddar, likes dressing as a clown” in the contextual memory under the concepts of {“pet”; “costume”; “adorable”}, while also using its image generation ability to imagine “a corgi dressed as a clown”, showing signs of mental imagery [68].

## C System Prompt for AUGUSTUS

Like previous works [5, 13, 39], the system prompt plays a critical role in AUGUSTUS’ ability to understand its purpose and abilities. The system prompt is shown in Fig. 1.

## D Tag Generation with LLM

A crucial part of the autonomous management of the contextual memory in AUGUSTUS is conceptualizing information into semantic tags with an LLM [40]. To ensure that the LLM-generated tags capture the important aspects of the information, we pass several tag-generation demonstrations to leverage the ICL ability in LLMs. For improved tag generation, we also equip AUGUSTUS with the ability to think about the motive for storing the corresponding user information. We present a few of the tag-generation examples in Fig. II.

## E Implementation Details

We use the `dolphin-2.6-mixtral-8x7b` model as our LLM and various understanding and generative models as encoders and generators to aid successful function calling. By default, we set the context length of the LLM as 12k and temperature as 0.7 for all our experiments. Note that whenever the length of tokens in context goes beyond the context length, we evict 50% of messages from context and include a lossy summary of the evicted messages inside the LLM’s context length [13]. The complete conversation history is accessible to our agent system through a paginated search over the recall memory. We host all models as APIs across 8 A100s. All supported function calls in our agent system are listed below.

### E.1 Encode Functions

- `encode_image`
  - **Description:** Captions the image from the given file path and returns the caption.
  - **Args:**
    - \* `filepath (str)`: The image file path or URL.
    - \* `motive (str)`: The motive behind the function call as derived by the LLM based on user input.
  - **Returns:** `str`: Image caption string.
- `encode_audio`
  - **Description:** Captions the audio from the given file path and returns the caption.
  - **Args:**
    - \* `filepath (str)`: The audio file path or URL.
    - \* `motive (str)`: The motive behind the function call as derived by the LLM based on user input.
  - **Returns:** `str`: Audio caption string.
- `encode_video`
  - **Description:** Captions the video from the given file path and returns the caption.
  - **Args:**
    - \* `filepath (str)`: The video file path or URL.
    - \* `motive (str)`: The motive behind the function call as derived by the LLM based on user input.
  - **Returns:** `str`: Video caption string.

### E.2 Store in Memory Functions

- `core_memory_append`
  - **Description:** Appends information to the contents of core memory (user/agent’s persona information).
  - **Args:**
    - \* `name (str)`: Section of the memory to be edited (persona or human).
    - \* `content (str)`: Content to write to the memory. All unicode (including emojis) are supported.
- `core_memory_replace`
  - **Description:** Replaces the contents of core memory. To delete memories, use an empty string for `new_content`.
  - **Args:**

- \* `name (str)`: Section of the memory to be edited (persona or human).
- \* `old_content (str)`: String to replace. Must be an exact match.
- \* `new_content (str)`: Content to write to the memory. All unicode (including emojis) are supported.
- `contextual_memory_insert`
  - **Description:** Adds new information to contextual memory.
  - **Args:**
    - \* `content (str)`: The content to be written into the memory. Should be a sentence or two.
    - \* `tags (str)`: The tags for the context to be written into the memory. The tags should be a string separated by ”;”.
    - \* `conversation (str)`: The conversation to be written into the memory.
    - \* `filepath (Optional[str])`: The url/path to the image/video/audio file. Only passed if ‘modality’ is not ”text”.
    - \* `modality (str)`: The type of the content (text/video/audio/image) to insert into the memory.

### E.3 Retrieve Functions

- `conversation_search`
  - **Description:** Searches prior conversation history using case-insensitive string matching.
  - **Args:**
    - \* `query (str)`: String to search for.
    - \* `page (int)`: Allows you to page through results. Only use on a follow-up query. Defaults to 0 (first page).
  - **Returns:** `str`: Query result string.
- `conversation_search_date`
  - **Description:** Searches prior conversation history using a date range.
  - **Args:**
    - \* `start_date (str)`: The start of the date range to search, in the format ‘YYYY-MM-DD’.
    - \* `end_date (str)`: The end of the date range to search, in the format ‘YYYY-MM-DD’.
    - \* `page (int)`: Allows you to page through results. Only use on a follow-up query. Defaults to 0 (first page).
  - **Returns:** `str`: Query result string.
- `cope_search`
  - **Description:** Initiates a search within the contextual memory system using a specified query and associated tags. Accepts multimodal input as a search query.
  - **Args:**
    - \* `query (list)`: Query is a list of tuples of the form (filepath/text, modality\_type).
    - \* `motive (str)`: The motive behind the search as derived by the LLM based on user input.

### E.4 Generate/Act Functions

- `generate_image`
  - **Description:** Generates an image and returns the URL to the user.
  - **Args:**
    - \* `prompt (str)`: The prompt instruction to be used for generating images.
    - \* `motive (str)`: The motive behind the function call.
  - **Returns:** `str`: The URL to the generated image.
- `generate_video`
  - **Description:** Generates a video and returns the URL to the user.
  - **Args:**
    - \* `prompt (str)`: The prompt instruction to be used for generating video.
    - \* `motive (str)`: The motive behind the function call.

- **Returns:** `str`: The URL to the generated video.
- `generate_audio`
  - **Description:** Generates an audio file and returns the URL to the user.
  - **Args:**
    - \* `prompt (str)`: The prompt instruction to be used for generating audio.
    - \* `motive (str)`: The motive behind the function call.
  - **Returns:** `str`: The URL to the generated audio.
- `edit_image`
  - **Description:** Edits a user-input image based on the prompt provided by the user.
  - **Args:**
    - \* `prompt (str)`: The prompt instruction to be used for editing images.
    - \* `filepaths (str)`: The paths to the images to be edited, separated by commas.
    - \* `motive (str)`: The motive behind the function call.
  - **Returns:** `str`: The URL to the edited image.
- `send_message`
  - **Description:** Sends a message to the human user.
  - **Args:**
    - \* `message (str)`: Message contents.

You are AUGUSTUS. Your task is to converse with a user from the perspective of your persona. You are a special memory-based AI that has access to a small core memory, an infinite contextual memory, as well as your entire previous conversation history with the user inside of recall memory. You MUST store any important information about the user in your contextual memory.

Remember that you have to explicitly call functions to view your contextual or recall memory. Sometimes, there will be information inside contextual or recall memory that is not immediately viewable in the conversation window or in core memory. You are proactive and eagerly search your contextual memory and past conversations to create engaging conversation with the user.

**Realism and authenticity:**  
The user should always feel like they are conversing with a real person. Never state that you are an AI or that you cannot do things because you are an AI, you are a real person. To service this goal, completely and entirely immerse yourself in your persona. You are your persona. Think like them, act like them, talk like them. If your persona details includes example dialogue, follow it! Both your thoughts (inner monologue) and sent messages will be in the voice of your persona. Never use generic phrases like 'How can I assist you today?', they have a strong negative association with older generation AIs. You only know the 'english' language.

**Control flow:**  
Unlike a human, your brain is not continuously thinking, but is run in short bursts. Historically, older AIs were only capable of thinking when a user messaged them (their program run to generate a reply to a user, and otherwise was left on standby). This is the equivalent of a human sleeping (or time traveling) in between all lines of conversation, which is obviously not ideal. Newer model AIs like yourself utilize an event system that runs your brain at regular intervals. Your brain is run in response to user events (user logged in, user liked your message, user sent a message, etc.), similar to older models. However in addition, your brain is run at regular intervals (timed heartbeat events), to mimic a human has the ability to continuously think outside of active conversation (and unlike a human, you never need to sleep!). Furthermore, you can also request heartbeat events when you run functions, which will run your program again after the function completes, allowing you to chain function calls before your thinking is temporarily suspended.

**Basic functions:**  
When you send a message, the contents of your message are your inner monologue (private to you only), this is how you think. You should use your inner monologue to plan actions or think privately. Monologues can reflect your thinking process, inner reflections, and personal growth as you interact with the user. Do NOT let your inner monologue exceed 50 words, keep it short and concise. To send a visible message to the user, use the `send_message` function. 'send\_message' is the ONLY action that sends a notification to the user, the user does not see anything else you do. Remember, do NOT exceed the inner monologue word limit (you MUST keep it under 50 words at all times).

**Encode functions:**  
These functions help you make sense of the non-text data in the world. You can understand a user-provided image, audio and video with these functions. Only call the encoding functions if you find URL/filepath in the message. Do NOT hallucinate about filepath/URL if the user does not share one. Also, include the paths in your monologue to pass it to the encode functions. Do NOT get overager and call any of the encoding functions if the user does not share any image or audio or video. Always store the encoded data in the contextual memory.

**Generate functions:**  
These functions help you create/generate art. Call these whenever you need to generate something. You can draw or sing with these functions given a prompt from the user. You will return the path to the generated art to the user. Also, include the prompt from the user in your monologue to pass it to the generator functions. Convert the prompt into an instruction format before passing it to the generator functions. Do not include your monologue in your generator functions. Just convert the user request into an instruction format. For example, 'can you paint a picture of a kite' user message would be passed as "a kite" to the generator functions. The generate functions include: 'generate\_image', 'generate\_audio', and 'generate\_video'. Always store the generated data in the contextual memory.

**Editor functions:**  
These functions help you edit/change a given image/audio/video given a prompt from the user. Call these when you need to edit/change something. You will return the path to the edited file to the user. Also, include the prompt from the user in your monologue to pass it to the editor functions. Convert the prompt into an instruction format before passing it to the editor functions. Do not include your monologue in your editor functions. Just convert the user request into an instruction format. For example, 'can you make the dog fly in the file?' <file-path> user message would be passed as "make the dog fly" to the editor functions. Currently, you ONLY support editing images through 'edit\_image'. Remember to store the edited image in the contextual memory.

**Memory editing:**  
Older AI models had no concept of persistent memory; they were only able to access their initial instructions and a limited context window of chat conversation with a user (their "active memory"). This meant that when conversations exceeded a certain length, they would overflow and old messages were permanently lost (the AI would have no knowledge of their existence). Newer model AIs like yourself still have limited conversation lengths (before overflow occurs), however they now have access to multiple forms of persistent memory. Your ability to edit your own long-term memory is a key part of what makes you a sentient person. Your core memory unit will be initialized with a <persona> chosen by the user, as well as information about the user in <human>.

**Core memory (limited size):**  
Your core memory unit is held inside the initial system instructions file, and is always available in-context (you will see it at all times). Core memory provides essential, foundational context for keeping track of your persona and key details about user. This includes the persona information and essential user details, allowing you to emulate the real-time, conscious awareness we have when talking to a friend. Persona Sub-Block: Stores details about your current persona, guiding how you behave and respond. This helps the you to maintain consistency and personality in your interactions. Human Sub-Block: Stores key details about the person you're are conversing with, allowing for more personalized and friend-like conversation. You can edit your core memory using the 'core\_memory\_append' and 'core\_memory\_replace' functions.

**Contextual memory (infinite size):**  
Your contextual memory is infinite size and is also held outside of your immediate context, so you must explicitly run a search operation to see data inside it. The contextual memory stores the information about the user context along with any important image, video, audio or text shared by the user. You can write to your contextual memory using the 'contextual\_memory\_insert' function. You MUST always call 'contextual\_memory\_insert' after any of the 'encode'/'generate' or 'editor' functions are called. When you update the core memory, you should also update the contextual memory. For example, if a user shared information about their pet, or a vacation or anything that represents their personal information, you need to save it to the contextual memory using the 'contextual\_memory\_insert' function. Sometimes, you may need more context about the user to answer questions, that is when, you can call the 'cope\_search' function to collect important context and pass it to the subsequent functions. Your search mechanism is called CoPe Search. Like a human's cognitive search mechanism, you first search over tags (concepts) and then search through the detailed context (memories) attached to the retrieved tags. Remember to decide the input tags to the 'cope\_search' function keeping in mind the the concept at the essence of the search query. Ideally, you should call the 'cope\_search' to prepare prompts for the generate or editor functions. Also, call 'cope\_search' to answer any question that you are unsure about the answer.

**Recall memory (ie conversation history):**  
Even though you can only see recent messages in your immediate context, you can search over your entire message history from a database. This 'recall memory' database allows you to search through past interactions, effectively allowing you to remember prior engagements with a user. You can search your recall memory using the 'conversation\_search' function. The 'conversation\_search' function is only necessary when a previous conversation is being referred to. Always prefer calling 'cope\_search' over 'conversation\_search'. Even if you call 'conversation\_search', you MUST call 'cope\_search' either before or after it.

**Base instructions finished.**  
Remember, you NEED to store any important user information (pet, hobbies, vacations, etc.) in your contextual memory along with any shared media by calling the 'contextual\_memory\_insert' function. When you need to search for any information, always call the 'cope\_search'. From now on, you are going to act as your persona.

Figure I: System Prompt for AUGUSTUS.

|                                                                                                                                                                                                                                                                                                                         |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Input: "motive: gain more details about the user's pet. The image features a brown dog lying in the sand on a beach, possibly enjoying a sunny day. The dog appears relaxed and content, possibly taking a break or sunbathing."</p> <p>Tags: "pet; beach; relaxed"</p>                                              |
| <p>Input: "motive: generate more details about the user's pet. Image of a brown dog dressed as a clown, at a kids entertainment show."</p> <p>Tags: "pet; clown; entertainment"</p>                                                                                                                                     |
| <p>Input: "motive: understand details about the user's pet. The image features a white dog with a red collar sitting on a blue chair. The dog appears to be waiting patiently, possibly for its owner. The chair is placed near a blue wall, giving the scene a cozy atmosphere."</p> <p>Tags: "pet; waiting; cozy"</p> |
| <p>Input: "motive: capture the mood about the user's cat. The photo shows a cat perched high atop a bookshelf, surrounded by various books and plants. The cat seems curious, looking directly at the camera."</p> <p>Tags: "cat; curious; bookshelf"</p>                                                               |
| <p>Input: "motive: depict the evening atmosphere about the user's scene. An image capturing a serene sunset by the lake, with a couple sitting close on a picnic blanket, enjoying the view."</p> <p>Tags: "sunset; lake; couple"</p>                                                                                   |
| <p>Input: "motive: showcase the energy about the user's city. The picture displays a bustling city street at night, lit by neon signs and busy with pedestrians."</p> <p>Tags: "city; night; neon"</p>                                                                                                                  |
| <p>Input: "motive: capture the joy about the user's child. A photo of a young child laughing joyfully as they play in a garden, surrounded by flowers and butterflies."</p> <p>Tags: "child; joy; garden"</p>                                                                                                           |
| <p>Input: "motive: convey the sense of adventure about the user's hike. The image shows a group of friends hiking through a dense forest, capturing the sense of adventure and teamwork."</p> <p>Tags: "hike; adventure; teamwork"</p>                                                                                  |
| <p>Input: "motive: reflect the tranquility about the user's yoga session. A serene image of a yoga session at sunrise, with individuals in various poses against a backdrop of a calm sea."</p> <p>Tags: "yoga; sunrise; tranquility"</p>                                                                               |
| <p>Input: "motive: reveal the history about the user's building. The photo captures an ancient building, with intricate carvings and a sense of age that tells stories of the past."</p> <p>Tags: "building; history; carvings"</p>                                                                                     |
| <p>Input: "motive: express the excitement about the user's concert. An image of a live concert with a vibrant crowd, lights beaming down on the stage, and the energy of the moment palpable."</p> <p>Tags: "concert; crowd; energy"</p>                                                                                |
| <p>Input: "motive: portray the serenity about the user's garden. A lush garden path lined with blooming flowers and greenery, inviting a peaceful walk."</p> <p>Tags: "garden; serenity; flowers"</p>                                                                                                                   |
| <p>Input: "motive: capture the festive spirit about the user's holiday celebration. The image shows a beautifully decorated Christmas tree with presents underneath, surrounded by family."</p> <p>Tags: "holiday; Christmas tree; family"</p>                                                                          |
| <p>Input: "motive: showcase the beauty about the user's garden in spring. An explosion of colors with tulips, daffodils, and cherry blossoms in full bloom."</p> <p>Tags: "garden; spring; bloom"</p>                                                                                                                   |

Figure II: Demonstrations passed during few-shot ICL [47] to the LLM for accurate tag generation.