Not all images are worth a thousand words!

Koustav Banerjee, Hardik Gupta, Lin Xie
Visual Linguists
baner188@umn.edu, gupt0414@umn.edu, lanca065@umn.edu

Abstract

Human memory for abstract visuals is notably poorer than for real-world scenes. We investigated whether this discrepancy stems from the difficulty in verbally labeling abstract content. We utilize a set of abstract images that vary in their complexities and human memory performance on those images to test our hypothesis. Using LLMs to generate captions for these abstract images of varying complexity, we found that caption complexity mirrored visual complexity. We built several simple models to predict human memory performance on abstract images and found that a predictive model based on the generated captions' semantic content performed competitively to a model relying solely on visual features. This study suggests that our ability to verbalize and name what we are seeing might indeed underpin our ability to remember a visual scene.

1 Introduction

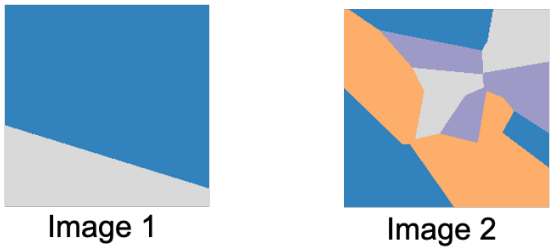

Figure 1: Representative images from our dataset

It can be easily agreed that image 2 in Figure 1 is harder to memorize than Image 1. While both these images are of the same spatial resolution, we can spot that Image 2 is more complex visually. Previous work has shown that this visual complexity causes an image to be harder to memorize (Banerjee, 2024). In this work we explore whether the ability to describe the image also has an effect on our ability to memorize the image. In other words, our visual memory might be aided by the ability to semantically represent the image. We characterize what makes an image harder to memorize by considering the hypothesis: as the images become more abstract, they become harder to describe and semantically represent, and that causes them to be harder to memorize.

So far, the visual memory literature has mostly considered visual perceptual cues to indicate difficulty in memorization. As a result we have got very well-established theories that we have a 'magic number'*4-object limit memory* (Luck and Vogel, 1997) that has been tested on multiple occasions (Ngiam et al., 2023). However, we can recall a large amount of information about natural scenes (Brady et al., 2016), which seems to go against these well established hypothesis. We investigate the reason for the hard limit of memorizing artificial stimuli in this work which does not seem to exist for natural scenes. [1]

2 Background

Intuitively, as when we read a book, we can visualize the scene the words describe indicating that linguistic stimuli can generate visual imagery. In this work, we check for the opposite effect, investigating if visual stimuli would cause linguistic representations.This approach has been utilized in recent computational domains to reconstruct images from fMRI signals (Liu and et al., 2023). However, this approach has not been used to actually explain human behavior, to the best of our knowledge. If the linguistic representation of the images occurs during visual perception, then as the image becomes more complex, we expect a positive correlation to the linguistic description of the image. With the advent of modern frameworks like Transformers (Vaswani et al., 2017), natural language processing has seen a massive boom since the first Large-Language Model (LLM) (Devlin et al., 2018). In

[1] Find our project website using this link

this work, we have utilized GPT-4 (Achiam et al., 2023) to generate descriptions of the images. We have used these generated descriptions to see if descriptions get more complex as the images get more abstract, and if that explains human memory performance.

Although purely visual memory has been found that purely visual memory exits (Lin et al., 2021), we try to look at whether this memory is facilitated further by the addition of semantic labels. Semantic labels and representations significantly enhance visual memory by providing organizational frameworks that anchor visual stimuli to meaning. In cognitive psychology, the dual-coding theory posits that information encoded both visually and verbally is more easily remembered because the two representations create redundant pathways for retrieval (Paivio, 1991). Semantic labels enable individuals to form associative links between visual elements and their meanings, transforming raw perceptual input into structured, conceptually rich representations. For example, studies on memory for scenes have shown that the presence of semantic information, such as recognizable objects and contextual relationships, boosts memory retention by engaging linguistic and conceptual processing systems alongside visual pathways (Brady et al., 2008). This interplay between visual and semantic encoding allows for efficient chunking of information, reducing cognitive load and enhancing recall accuracy.

Neuroscience studies further corroborate the role of semantic representations in visual memory by highlighting the involvement of brain regions like the medial temporal lobe and the prefrontal cortex, which are associated with integrating sensory input and conceptual knowledge. Semantic associations activate higher-order brain networks that enhance the distinctiveness of visual stimuli, making them more memorable. Binder and Desai (2011) showed that semantic memory relies on distributed neural systems, with hubs like the angular gyrus contributing to the processing of conceptual information (Binder and Desai, 2011). This is crucial for understanding and remembering complex scenes. By providing a cognitive scaffold, semantic labels and representations enable humans to organize and retrieve visual experiences with remarkable efficiency, particularly in naturalistic environments rich in recognizable features and meaningful context.

Thus it makes sense to study how semantic representation and describability of images may aid visual memory. Advancements in image captioning have shown significant promise in enhancing our understanding of visual-semantic interactions. For instance, FUSECAP utilizes vision-language pretraining and "vision experts" (e.g., object detectors and OCR systems) to generate enriched captions by integrating detailed visual features into existing captions (Rotstein et al., 2023). These enriched captions align better with image content and improve tasks such as image-to-text retrieval, showcasing the potential of dataset-focused improvements for enhancing caption quality.

Another approach, CapText, avoids direct image processing by generating captions from textual descriptions and context alone, leveraging large language models (LLMs) (Ghosh and Anupam, 2023). This computationally efficient method aligns well with the challenges of captioning abstract images, where context often plays a more significant role than visual features. However, these methods have not been explicitly tailored to abstract stimuli, leaving a gap in understanding their applicability to non-realistic imagery.

Furthermore, cognitive studies on LLMs reveal their emergent memory-like traits, such as primacy and recency effects, which mirror human memory processes (Janik, 2024). These traits suggest that LLMs could serve as tools to model human cognitive behaviors, particularly in scenarios where linguistic representations intersect with visual perception.

Despite these advancements, challenges persist. Research highlights the failure of existing captioning systems to produce meaningful descriptions for abstract images, often due to a reliance on literal interpretations (Tran et al., 2022). Testing frameworks like MetaIC further expose the limitations of current systems in handling synthetic visuals, emphasizing the need for improved robustness in abstract contexts (Authors, 2022).

This work aims to bridge the gap between computational and cognitive approaches by investigating the role of semantic complexity in image captioning and its relationship with human memory performance. By leveraging GPT-4 to generate captions for abstract images, we seek to understand how the linguistic complexity of captions correlates with memory challenges for abstract stimuli. This study contributes to both the computational modeling of memory processes and the development of more robust captioning systems for abstract visual

data. These LLM generated captions for images can be analyzed for their descriptive complexity, and compared to the images' visual complexity.

The complexity of a sentence can be calculated using various techniques, each focusing on different aspects of the sentence. Perplexity measures how well a statistical model predicts a sample. In the context of sentences, it's used to gauge the predictability of words in a sentence. A higher perplexity indicates a more complex sentence because it means the model finds it harder to predict the next word (Jelinek et al., 1977).

Syntactic Complexity (Lu and Ai, 2015) measures the complexity of sentence structure, including the number of clauses, phrases, and grammatical dependencies. Metrics like dependency distance (Liu et al., 2017) are commonly used to calculate this. Type-Token ratios can be used to calculate the diversity of vocabulary used in the captions (Kettunen, 2014).

Concreteness is another measure used to analyze sentence complexity and highly pertinent to our work (Begg and Paivio, 1969). Concreteness refers to the degree to which a word or concept is tangible and easily visualized. In NLP, concreteness is often used to assess the complexity of sentences by evaluating the abstractness of the vocabulary used. More abstract words typically make sentences harder to understand, while concrete words are easier to visualize and comprehend. Neuroscience and psychological works have shown that sentence concreteness has shown to capture ways people speak about concepts (Snefjella and Kuperman, 2015), and also individual differences among people's behaviors (Botch and Finn, 2024). Several methods of measuring concreteness has been proposed (Ljubešić et al., 2018), (Yanuka et al., 2024). These methods, however, use complex models to calculate the concreteness. Here, we have used word synsets to build a simple metric for concreteness.

The potential impact of this work is substantial. By demonstrating a correlation between semantic complexity and memory performance, this research could guide the development of more effective educational tools, where abstract visual content is paired with semantically enriched descriptions. Additionally, in practical applications like assistive technologies or augmented reality systems, better understanding of how humans interpret abstract images can lead to improved user interfaces and interactions. Finally, this study could pave the way for future innovations in multimodal AI systems that emulate human-like memory for complex visual stimuli.

3 Methodology

We employed a multi-step approach to investigate the influence of semantic complexity and concreteness in image captions on human memory performance.

Caption Generation

Image captions were generated using the gpt-4o-mini language model. Captions were designed through prompt engineering to ensure relevance and detail, focusing on spatial relationships and object shapes (see Figure 2 (Prompt Selection)). Multiple such prompts were used to generate the captions. These prompts were used to generate captions for a single block of abstract images.

These captions generated from the different prompts were compared to whittle down and select the final prompt. We evaluated captions using both automatic (ROUGE-1 scores) and human (normalized Likert scale ratings) metrics. These ratings showed a clear preference for our most tuned prompt (see Figure 3).

The final prompt that we selected and used to generate captions for the entire dataset was:

"Please generate descriptions of the images and talk about the spatial relationships and the object shapes of all the colors in the image in detail without using poetic or emotional language. Write a maximum of 100 words."

Semantic Complexity

Once we had all the captions for the abstract images, our next step was to evaluate the semantic complexity of the captions for those images (See Figure 2 (Model Building section)). Semantic complexity was quantified using **perplexity** (Jelinek et al., 1977) and **semantic concreteness**(Begg and Paivio, 1969), which were calculated from a pretrained BERT model (bert-base-uncased) and using Wordnet(Fellbaum, 1998), respectively. Perplexity was computed by tokenizing captions, passing them through the model, and taking the exponential of the cross-entropy loss.

Concreteness Calculation

To assess concreteness, tokens were mapped to WordNet synsets, and synset depths were used to compute concreteness scores. Caption-level scores

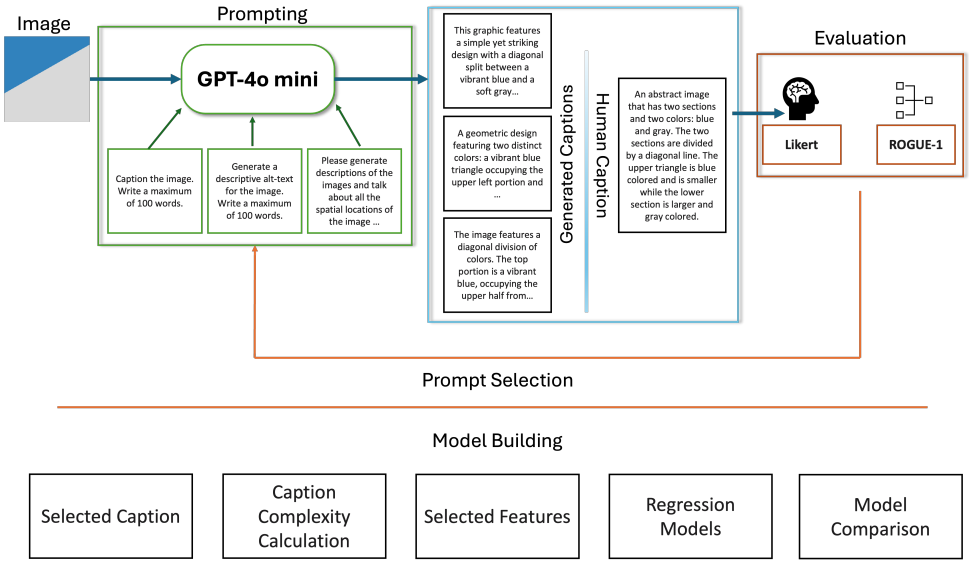

Figure 2: Methodology

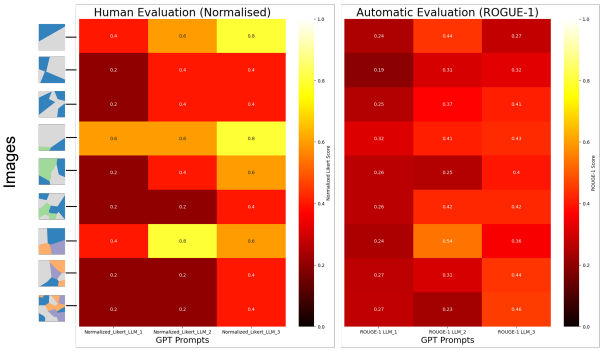

Figure 3: Human and ROGUE-1 evaluations for captions generated different prompts for 1 block of abstract images.

were averaged across all tokens. While other methods to calculate concreteness exist, we went ahead with this as it utilizes the well-established WordNet library.

Feature Selection

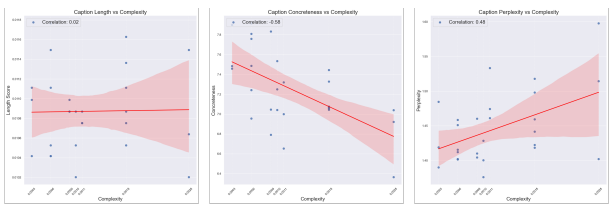

Figure 4: Different features extracted from captions are used as predictors for human memory performance. Here we check whether these semantic features are correlated to visual complexities of the images.

We computed Pearson's correlation coefficients between semantic complexity measures (perplexity, concreteness, and memorability) and distortion metrics, a proxy for human memory performance.

Regression Models

Linear regression models were trained using semantic complexity scores (perplexity, concreteness, caption length) as features (See Figure 4). A linear combination of perplexity and concreteness were used to create a semantic memorability model. This model aims to predict human memory performance on the visual task purely based on linguistic complexity and descriptive specificity. All models models were evaluated using K-fold cross-validation, by treating each block of images as a fold (see Dataset section for details), to ensure independence between training and testing sets.

4 Challenges

One of the primary challenges was that existing large language models (LLMs) haven't been trained extensively on abstract images. These models generally excel at interpreting and describing real-world, concrete scenes, objects, and events because they have been exposed to vast amounts of such data. However abstract images, which are often used as controlled stimuli for scientific research and lack objects encountered in the everyday world, are underrepresented in typical training datasets.

As a result, when tasked with generating captions for abstract images, these models tend to fall short, producing captions that fail to describe the image accurately. The initial generated captions defaulted to a naturalistic interpretation of the ab-

stract images. For instance, Figure 1's blue and white image was captioned as *"glacier meeting the sea"* which is not an expected response from a human. This gap highlights a significant limitation in current LLM capabilities and underscores the need for more diverse and comprehensive training datasets that include a variety of abstract and non-natural stimuli.

To mitigate this issue, we had to employ thorough prompt engineering. This involved crafting prompts that would guide the model towards responding how human may have responded while performing a scientific task. As a result, we had to manually test and validate multiple prompts. This labor-intensive process was reminiscent of the early days of creating image datasets, where human annotators painstakingly labeled each image. This step was essential not only for quality control but also for refining our approach. In the end, this step enabled us to successfully devise a caption-generation technique for abstract images.

We encountered other challenges as well such as having limited measures of analyzing the complexity of sentences. Readability metrics such as Flesch Reading Ease (Farr et al., 1951) and Flesch-Kincaid scores (Flesch, 2007) were tested but excluded due to poor predictive performance. Other limitations can be thought of as our dataset being limited to 27 images and participants being limited to 6 participants. While 27 images is standard for a psychological task such as ours, we can look into getting memory performances of more participants in the future.

5 Results

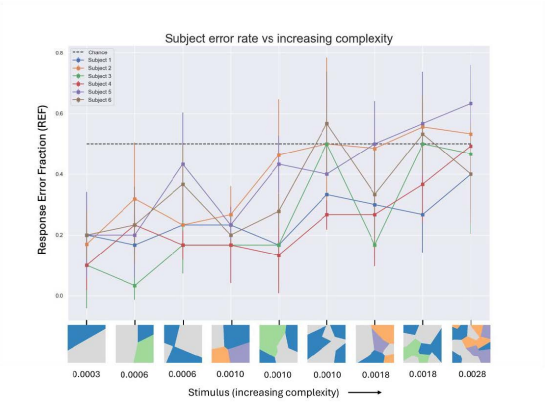

Figure 5: Human memory performance shows increasing errors with increasing image complexity. *Image and caption adapted with permission from author(Banerjee, 2024).*

Recent work has shown that as image complexity increases, human memory performance declines for abstract images (See Figure 5).

A UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) of the captions of select group of images was used to cluster captions generated for abstract images by their caption complexity. The UMAP demonstrates clear separation among captions based on their associated image complexities (See Figure 6). This separation indicates that the captions, reflect underlying patterns of image complexity that humans intuitively perceive.

Figure 6: UMAP of captions for selected images show clustering by image complexities.The image groups were chosen such that all the images in a group had the same visual complexity, but the groups were of different complexities. The group visual complexities: 0.0003, 0.0010, 0.0028 (for more details on the images see 8)

We also compared the performance of models predicting memory performance using different semantic complexity metrics (See Figure 7). A K-Fold cross-validation approach was employed to rigorously compare a **semantic memorability model** (based on a linear combination of perplexity and concreteness) with a purely visual complexity model. Remarkably, the semantic model proved competitive with the visual complexity model, underscoring the importance of semantic factors in determining memorability.

6 Additional Points

Replicability

Whilst our results are tied to the dataset that we used, our methodology is designed to be highly replicable. We used publicly available tools, such as GPT-4 and pre-trained BERT models, ensuring that others can recreate our process. Additionally, our prompts for caption generation are explicitly

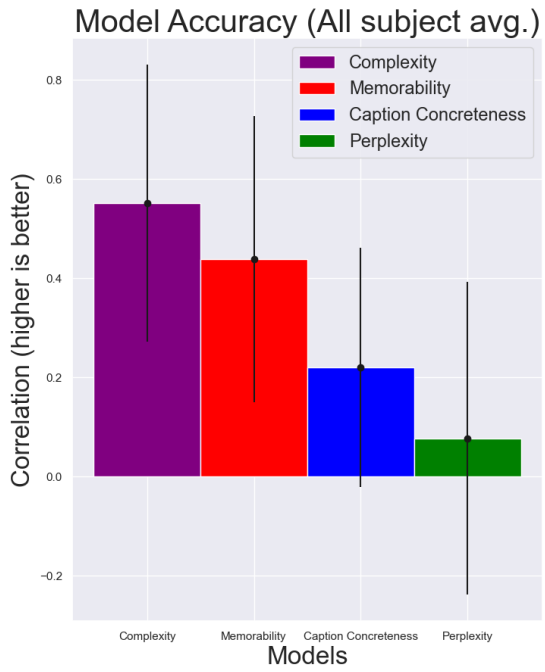

Figure 7: A K-Fold cross-validation approach to compare models shows that semantic memorability model is competitive with a visual complexity regression model. This gives evidence to the hypotheis that images may be represented semantically in addition to being represented visually.

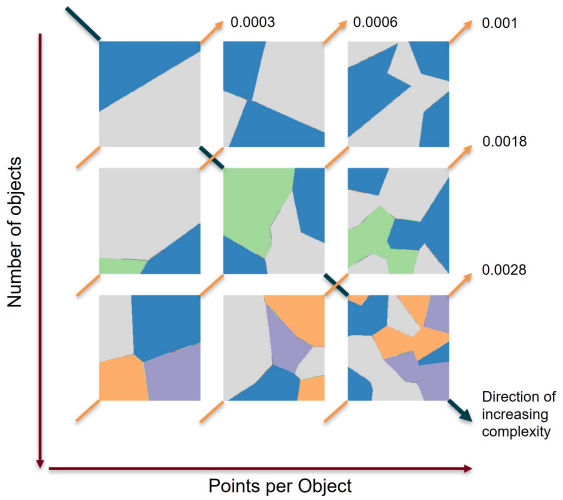

Figure 8: One block of the stimuli that participants saw during the memory task. It consists of 3x3 images, with increasing number of objects as we go down, and increasing fragmentation of objects as we go across a row. Numbers over the arrows indicate the visual complexites of the images. The visual complexity of each images, is a unitless metric derived from structural and geometric properties of the image. Orange arrows signify images with same visual complexities. *Image and caption adapted with permission from author(Banerjee, 2024).*

detailed, allowing for straightforward replication. Sharing our annotated dataset and codebase will facilitate broader use and validation of our findings.

Datasets

We used a recently created custom dataset for a mental imagery and human memory task (Banerjee, 2024). The dataset consists of 3 blocks of non-natural images with varying number of objects, fragmentation patterns, and complexities, with each block having 9 images (See Figure 8). This dataset containing the abstract images was paired with with GPT-4-generated captions for this task. The combination of abstract images and robust captions provides a unique resource for future research. The dataset may prompt researchers to explore novel approaches to abstract image captioning, such as developing models specifically tailored for non-naturalistic visual stimuli.

Ethics

We ensure that our prompts are designed to avoid bias or emotionally charged language.

Discussion

Overall, the results strongly support the idea that semantic descriptions aid visual memory. They reveal a robust connection between image complexity, semantic representation, and memory performance, shedding light on why natural scenes are easier to remember than abstract ones. The implications of this study extend to areas such as AI-driven image captioning, memory training, and understanding the cognitive basis of visual and semantic integration.

In this study, we quantitatively analyzed human visual memory performance through semantic means by leveraging metrics such as perplexity and concreteness. These metrics serve as proxies for the difficulty of generating semantic descriptions, revealing that abstract images with higher visual complexity lack clear semantic cues, leading to increased memory recall errors. This supports the hypothesis that humans rely on both visual and semantic representations to encode and retrieve memories, with the absence of semantic anchors in abstract images impairing memory performance.

Our UMAP analysis demonstrated that captions for less complex images cluster tightly, reflecting

straightforward semantic structures, while captions for more complex images are dispersed, indicating greater variability and difficulty in interpretation. This suggests that semantic properties are closely tied to visual complexity, offering new insights into how image semantics influence cognitive processes.

By combining semantic and visual complexity metrics, we achieved more accurate predictions of memory performance. This underscores the dual role of semantic and visual encoding in memory retention and highlights the importance of integrating these perspectives for a holistic understanding of human cognition. Our work serves as a proof of concept, illustrating that harder-to-describe images are also harder to memorize, while bridging two fields to explore the underexamined relationship between human memory and semantic representation. Furthermore, our approach demonstrates the utility of leveraging LLMs for generating captions for abstract images, opening avenues for future research and controlled stimuli development.

We believe, that this work will interest a wide range of audiences, including cognitive psychologists, computational neuroscientists, and artificial intelligence researchers, particularly those working at the intersection of human cognition and machine learning. Cognitive psychologists can utilize the findings to deepen their understanding of how humans process and recall abstract visual information. Computational neuroscientists might find value in the methodology, especially the integration of semantic and linguistic complexity to model memory. For AI researchers, this study offers insights into how large language models can simulate human cognitive processes, potentially improving systems designed for abstract image captioning.

7 Role assignment

1. Koustav Banerjee (Project design, coding, dataset labeling, writing)
2. Hardik Gupta (Website creation, writing, dataset labeling)
3. Lin Xie (Literature Survey)

References

Josh Achiam et al. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774. ArXiv preprint arXiv:2303.08774.

Anonymous Authors. 2022. Automated testing of image captioning systems. *arXiv preprint arXiv:2206.06550*.

Koustav Banerjee. 2024. *Evidence of Compression and Probabilistic Representation During Mental Imagery*. Ph.D. thesis, ProQuest Dissertations & Theses. ISBN: 9798342714860.

Ian Begg and Allan Paivio. 1969. Concreteness and imagery in sentence meaning. *Journal of Verbal Learning and Verbal Behavior*, 8(6):821–827.

Jeffrey R Binder and Rutvik H Desai. 2011. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536.

Thomas L Botch and Emily S Finn. 2024. Neural representations of concreteness and concrete concepts are specific to the individual. *Journal of Neuroscience*, 44(45).

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329.

Timothy F. Brady, Viola S. Stånormer, and George A. Alvarez. 2016. Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, 113(27):7459–7464.

Jacob Devlin et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805. ArXiv preprint arXiv:1810.04805.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Shinjini Ghosh and Sagnik Anupam. 2023. Captext: Large language model-based caption generation from image context and description. *Massachusetts Institute of Technology*.

Romuald A. Janik. 2024. Aspects of human memory and large language models. *Institute of Theoretical Physics and Mark Kac Center for Complex Systems Research, Jagiellonian University*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Qi Lin, Sami R Yousif, Marvin M Chun, and Brian J Scholl. 2021. Visual memorability in the absence of semantic content. *Cognition*, 212:104714.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Yulong Liu and et al. 2023. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. https://arxiv.org/abs/2302.12971. ArXiv preprint arXiv:2302.12971.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. *arXiv preprint arXiv:1807.02903*.

Xiaofei Lu and Haiyang Ai. 2015. Syntactic complexity in college-level english writing: Differences among writers with diverse l1 backgrounds. *Journal of second language writing*, 29:16–27.

Steven J. Luck and Edward K. Vogel. 1997. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

William XQ Ngiam, Krystian B. Loetscher, and Edward Awh. 2023. Object-based encoding constrains storage in visual working memory. *Journal of Experimental Psychology: General*.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2023. Fusecap: Leveraging large language models for enriched fused image captions. *Technion - Israel Institute of Technology*.

Bryor Snefjella and Victor Kuperman. 2015. Concreteness and psychological distance in natural language use. *Psychological science*, 26(9):1449–1460.

Tran et al. 2022. Bright as the sun: In-depth analysis of imagination-driven image captioning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Ashish Vaswani et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Moran Yanuka, Morris Alper, Hadar Averbuch-Elor, and Raja Giryes. 2024. Icc: Quantifying image caption concreteness for multimodal dataset curation. *arXiv preprint arXiv:2403.01306*.