Dimensions of Vulnerability in Visual Working Memory: An AI-Driven Approach to Perceptual Comparison

Yuang Cao (12310901@mail.sustech.edu.cn)[*]
Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China

Jiachen Zou (12210153@mail.sustech.edu.cn)[*]
Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China

Chen Wei (12150103@mail.sustech.edu.cn)[†]
Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China
Department of Psychology, University of Birmingham, Birmingham, United Kingdom

Quanying Liu (liuqy@sustech.edu.cn)[†]
Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China

Abstract

Human memory exhibits significant vulnerability in cognitive tasks and daily life. Comparisons between visual working memory and new perceptual input (e.g., during cognitive tasks) can lead to unintended memory distortions. Previous studies have reported systematic memory distortions after perceptual comparison, but understanding how perceptual comparison affects memory distortions in real-world objects remains a challenge. Furthermore, identifying what visual features contribute to memory vulnerability presents a novel research question. Here, we propose a novel AI-driven framework that generates naturalistic visual stimuli grounded in behaviorally relevant object dimensions to elicit similarity-induced memory biases. We use two types of stimuli—image wheels created through dimension editing and dimension wheels generated by dimension activation values—in three visual working memory (VWM) experiments. These experiments assess memory distortions under three conditions: no perceptual comparison, perceptual comparison with image wheels, and perceptual comparison with dimension wheels. The results show that similar dimensions, like similar images, can also induce memory distortions. Specifically, visual dimensions are more prone to distortion than semantic dimensions, indicating that the object dimensions of naturalistic visual stimuli play a significant role in the vulnerability of memory.

Keywords: Memory Distortion; Object Dimensions; AI-driven Generative Model

Introduction

Human visual working memory (VWM) serves as a critical cognitive interface, allowing us to temporarily retain and manipulate visual information to guide behavior (Luck & Vogel, 1997; Cowan, 2008; Vogel & Machizawa, 2004). Yet, this system exhibits marked vulnerability: even routine interactions with perceptual inputs—such as comparing a memorized object to a newly encountered one—can distort the original memory representation (Fukuda et al., 2022). Such distortions challenge the fidelity of both visual and semantic memory features, such as misremembering the color of an item, or an eyewitness misremembering whether a suspect was holding a weapon or a repair tool. While prior work has established that perceptual comparisons induce retroactive biases in VWM (termed similarity-induced memory biases, SIMB), three critical questions remain unresolved: (1) How to construct naturalistic visual stimuli that systematically elicit dimension-specific memory distortions? (2) Can perceptual comparisons involving similar abstract dimension activations, like those with similar images, also induce memory distortions? and (3) Do visual and semantic dimensions exhibit differential levels of susceptibility to memory distortions?

Existing research on memory distortion has predominantly relied on simplified stimuli (e.g., colors, shapes) to isolate mechanistic principles (Scotti, Hong, Leber, & Golomb, 2021; Chunharas, Rademaker, Brady, & Serences, 2022; Saito, Duncan, & Fukuda, 2023; Saito, Kolisnyk, & Fukuda, 2023; Saito, Bae, & Fukuda, 2024). Although these studies reveal that perceptual similarity amplifies memory biases, their conclusions are constrained by artificial experimental contexts. Real-world objects, in contrast, are defined by multidimensional and hierarchical attributes spanning low-level visual features (e.g., shape, texture) and high-level semantic properties (e.g., category, function). Recent evidence suggests that meaningfulness may enhance VWM capacity and stability (Asp, Störmer, & Brady, 2021; Sasin, Markov, & Fougnie, 2023), yet no study has compared the contribution of visual versus semantic dimensions to memory vulnerability. These gaps limit our understanding of how natural object representations interact with perceptual inputs to shape memory errors—a problem exacerbated by the lack of methods to precisely control and manipulate object dimensions in naturalistic stimuli.

The rise of AI-driven generative models offers a transformative solution. These models can synthesize naturalistic images, making them valuable tools for cognitive experiments (Karras, 2019; Wei, Zou, Heinke, & Liu, 2024b; Wei et al., 2025). However, their application to memory research remains nascent, particularly in the context of disentangling and editing latent object dimensions. Here, we bridge this gap by employing a novel AI-driven generative model that

[1] * Equal contribution
[2] † Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

generates natural visual stimuli through dimension manipulation (Wei, Zou, Heinke, & Liu, 2024a). This approach allows us to construct stimulus gradients along both holistic perceptual variations and isolated dimension activations—a capability essential for addressing our three core questions.

In this study, we examined memory vulnerability under two conditions: perceptual comparison with image wheels and perceptual comparison with dimension wheels (Fig. 1). Image wheels are generated by smoothly editing dimension activation values of an image, while dimension wheels are constructed from predefined activation values of dimensions in a latent space. These stimuli allow us to isolate the contributions of image similarity and dimension similarity to memory distortion during perceptual comparisons, as well as to disentangle the contributions of individual dimensions. We hypothesized that perceptual comparisons with both types of wheels would induce memory distortions, with visual dimensions exhibiting greater vulnerability than semantic dimensions. This hypothesis aligns with neurocognitive models positing that semantic features are better remembered thanks to schema-based representations and deeper processing (Brady, Störmer, & Alvarez, 2016; Brady & Störmer, 2022; Chung, Brady, & Störmer, 2023).

Our results confirm that similar object dimensions, whether manipulated via images or abstract dimension activations, can induce significant memory distortions. Additionally, visual dimensions showed markedly higher distortion susceptibility compared to semantic dimensions. These findings advance theoretical frameworks by demonstrating that memory vulnerability is not merely a function of perceptual similarity but is dimensionally structured. Methodologically, our AI-driven approach provides a blueprint for studying complex cognitive processes with naturalistic yet controlled stimuli, offering implications for AI-assisted experimental design and computational models of memory.

Related Works

Similarity-Induced Memory Bias. Studies have shown that when individuals compare a memorized stimulus to a similar perceptual input, the memory tends to shift towards the features of the new input, a phenomenon known as SIMB (Fukuda et al., 2022). In SIMB experiments, participants were asked to remember a target stimulus and later reproduce it by selecting from a continuous wheel of stimuli. When a pair of probe stimuli was introduced during the retention period, the memory of the target was significantly biased toward the probe judged to be similar (Saito, Duncan, & Fukuda, 2023). While SIMB has been observed with simple stimuli such as colors and shapes, its application to more complex, real-world objects remains understudied. Previous work primarily focused on low-level features, like color or shape (Fukuda et al., 2022; Saito, Duncan, & Fukuda, 2023; Saito, Kolisnyk, & Fukuda, 2023; Saito et al., 2024), but real-world objects are defined by multiple features (Hebart, Zheng, Pereira, & Baker, 2020; Bracci & Op de Beeck, 2023), including both low-level visual attributes (e.g., shape, texture) and high-level semantic properties (e.g., function, category). This suggests that evaluating SIMB with dimension-based sythetic stimuli, is essential to understanding how image and dimension similarity contribute to memory distortions, as well as the respective contribution of visual and semantic dimensions. This is particularly relevant when studying objects in their natural context, where dimensionality and feature diversity play significant roles in the perception and memory of objects.

Behavior-Based Object Dimensions. Over the past few years, behavioral-based object dimensions have attracted considerable attention, especially in computer vision and cognitive science research. Previous work (Hebart et al., 2019, 2020; Zheng, Pereira, Baker, & Hebart, 2019; Muttenthaler et al., 2022) has gathered human similarity judgment data for a naturalistic dataset of 26,107 object images through extensive behavioral experiments. These studies have decoded the representation dimensions of images by either optimizing object-specific representations or leveraging deep neural networks (DNNs) to predict human behavior, thus reducing complex visual information into interpretable, low-dimensional object features. Notably, recent findings (Kramer, Hebart, Baker, & Bainbridge, 2023) suggest that these dimensions can be used to construct an object feature model capable of predicting image memorability. Collectively, these studies highlight the value of behavior-based object dimensions in understanding visual perception and memory.

Dimension-Guided Wheel Generation

We employed the Concept-based Controllable Generation model from (Wei et al., 2024b, 2024a) to generate visual stimulus wheels. The model adopts a two-stage generation strategy: first modeling $p(h|e)$ and subsequently modeling $p(x|h)$, where e denotes conditioning parameters (dimension activation values or similarity), h represents CLIP embeddings, and x is the generated image. By incorporating training-free guidance during the modeling of $p(h|e)$, the model enhances generation flexibility and applicability. This framework allows effective guidance of visual stimulus generation through appropriate selection of conditioning parameters e and differentiable loss functions $\ell(f_\phi(\cdot),\cdot)$ based on experimental objectives.

For the generation of image wheels, we randomly select an image from the THINGS dataset and perform dimension editing on it. We used the following loss functions to guide the generation process:

1. *Dimensional Guidance*: We select two dimensions from the concept of the given image for editing. As shown in Fig. 1, we pre-designed the concept pair activation values for 12 images, arranging the activation values in a circular pattern within the representation plane defined by the concept pair axes. This guides the generation of a smooth transition of concepts across the image wheel. The corresponding loss

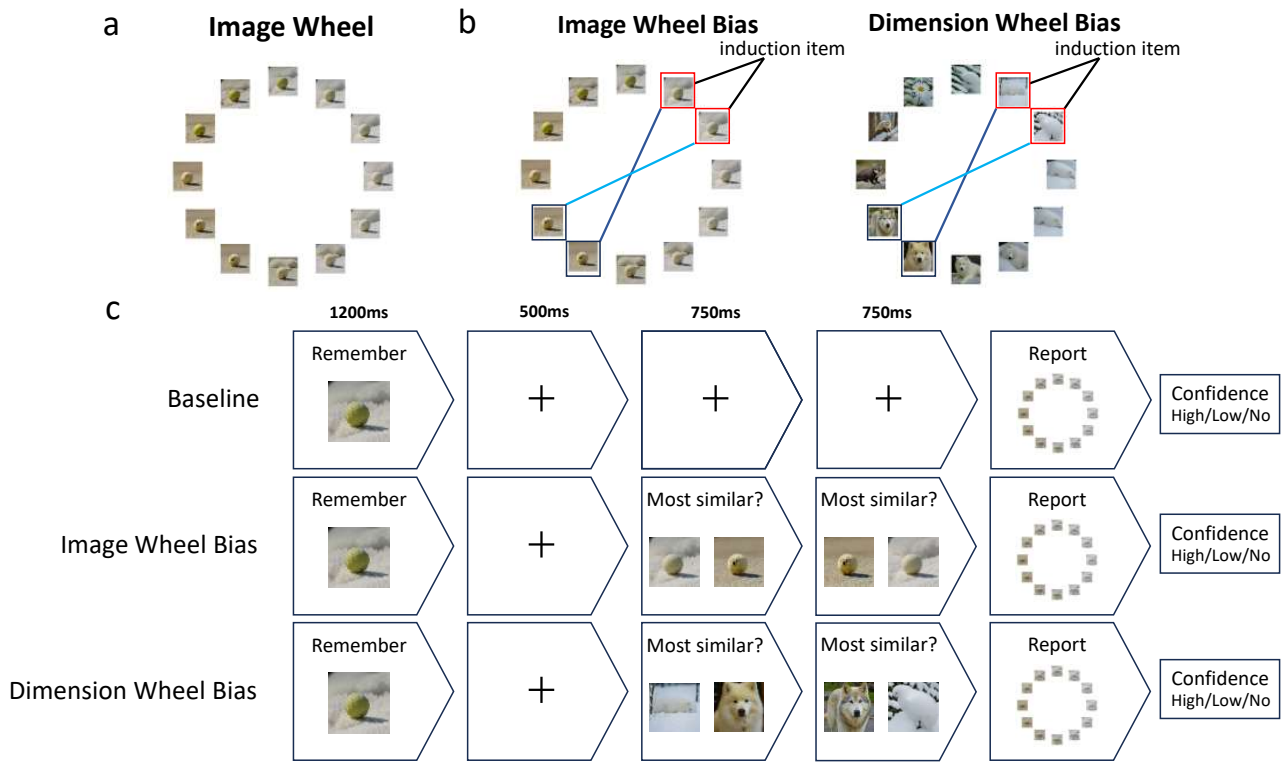

Figure 1: **Experimental Design**. (a) Image wheel: A circular arrangement of 12 images featuring gradual variation in characteristics. Participants were instructed to memorize the top image in the wheel (memory item). (b) Bias induction: Image wheel bias trials used a pair of images include a induction item (red box, selected clockwise from the memory item randomly) and another image (blue box) for similarity judgments, while dimension wheel bias trials employed pairs from dimension wheels, where images in dimension wheels were typically perceived as dissimilar to the memory item. (c) Experimental procedure: Three conditions (baseline, image wheel bias, dimension wheel bias). All conditions began with target memorization, followed by two similarity judgment phases (except baseline), and concluded with target recognition from the image wheel. Image pairs for similarity judgments were selected from image wheels in the image bias condition and from dimension wheels in the dimension bias condition.

function is:

$$\ell = \sum \|g(h)_i - c_i\| \tag{1}$$

where c_1-c_{12} constitute circular coordinates in the conceptual plane.

2. *Smoothness Guidance*: To ensure the images on the image wheel maintain similarity and smooth transitions, we implemented:

$$\ell = \sum_{i,j \in S} \|h_i - h_j\| \tag{2}$$

where $\{i, j\}$ denotes indices of neighboring images in the wheel.

3. *CLIP Guidance*: To preserve latent feature similarity with the given image, we used the following loss function:

$$\ell = \|h - \bar{h}\| \tag{3}$$

where $\bar{h}$ represents the CLIP embedding of the source image.

4. *Pixel Guidance*: We control the pixel-level similarity of the images by introducing img2img (Meng et al., 2021), which uses a noisy version of the given image as the starting point for generation.

For the generation of *dimension wheels* , we only applied dimension guidance. For each *image wheel*, we aimed to generate a dimension wheel by setting the predefined dimension pair activation values as the target. Since there are no restrictions on image similarity, the dimension wheels have lower visual similarity but maintain consistent dimensional similarity. This effectively separates the effects of visual similarity and dimension similarity on memory distortion.

Image Wheels Induction Experiment

Stimuli and Task Design

Image wheels were constructed using an AI-driven generative model that smoothly interpolated latent dimension activations to produce naturalistic object variations. Each wheel represented a circular manifold of images, where angular positions corresponded to incremental changes along predefined object dimensions. A memory item was randomly selected from an

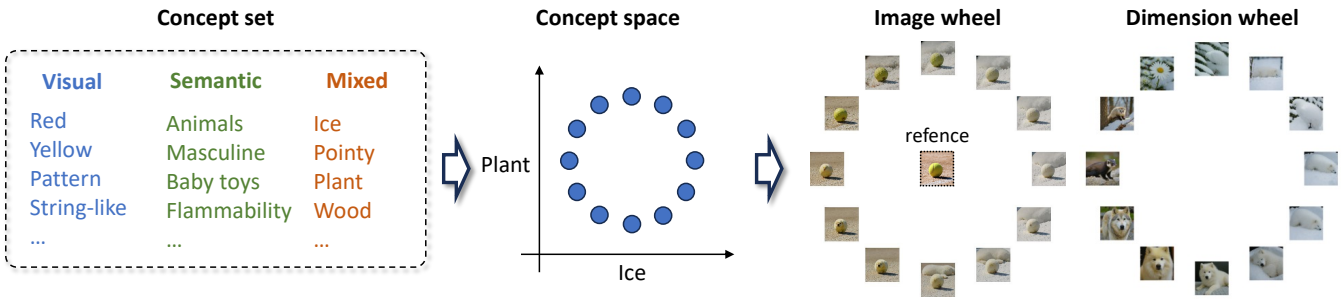

Figure 2: **Wheel generation.** The leftmost image shows the dimensions (with categories of visual, semantic, and mixed) that we selected during image generation. We created circles in the representation space of dimension-pairs as target dimension activation values, and then used an AI-driven generative model to generate wheels based on dimensions. For the image wheel, we generated wheels that are similar to the reference images but with different dimension activation values. For the dimension wheel, we used only the activation values in the dimension-pair space to generate wheels.

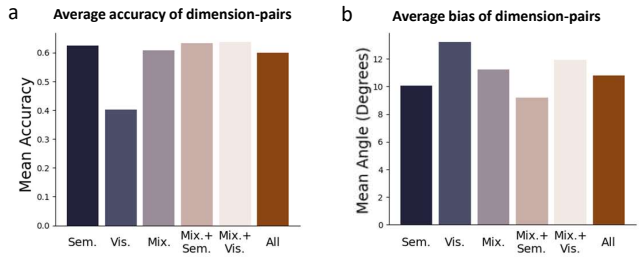

Figure 3: **Experimental results of the Image Wheels Induction Experiment (N=100).** (a) Mean accuracy of memory performance across dimension-pair categories. (b) Mean bias scores for dimension-pairs.
Sem. = Semantic, **Vis.** = Visual, **Mix.** = Mixed.

image wheel and presented for 1,200 ms. After a 1,500 ms retention interval, the picture wheel appeared, and the subject reported which picture in the wheel was the original memory item.

During the induction phase, two items were simultaneously displayed: an *induction item* and a *dissimilar item*. The induction item was sampled from a 60° arc along a target dimension direction relative to the memory item, while the dissimilar item occupied the position 180° opposite to the induction item. This design ensured that the induction item shared perceptual similarity with the memory item along the manipulated dimension, whereas the dissimilar item served as a control. Participants performed a same-different judgment task on these pairs for 1,500 ms. After a 500 ms retention interval, the picture wheel appeared, and the subject reported which picture in the wheel was the original memory item. The interstimulus interval is determined based on the results of the pre-test, in order to ensure that participants experience comparable task loads across different experimental conditions, thereby minimizing the potential interference of temporal variables on memory performance.

A total of 100 participants completed the experiment on the Brain Island platform. Each participant performed 120 trials, including 30 no-induction baseline trials and 90 induction memory trials. The experimental stimuli were categorized into three groups based on feature dimensions: *visual* (e.g., shape-color), *semantic* (e.g., tool-animal), and *mixed* (e.g., color-category). The mixed condition integrates both visual and semantic information dimensions, and we collected data from this condition to ensure the completeness of the study. However, the primary aim of this research was to compare the memory vulnerability differences between visual and semantic dimensions, so we did not conduct an in-depth analysis of the mixed condition. The results from the mixed condition were only used as supplementary information to help us gain a more comprehensive understanding of the potential impact of visual and semantic information integration on memory. The accuracy of the no-induction experiment reflects the participant's ability to remember the images. If a participant's overall accuracy in the no-induction experiment is below 40%, we do not accept their experimental results. In the induction experiment, the participant undergoes two induction trials. We consider the participant's final selection results in the image wheel to be valid only if they select the *induction item* in both induction trials, as this indicates their memory of the image is accurate. Additionally, at the end of each experimental trial, participants are asked to rate their confidence in their selection. We excluded low-confidence trials because they are less likely to reflect the participant's true memory and are more likely to be based on uncertainty or guessing. Excluding these trials helps ensure that the memory distortions we observe are due to actual memory effects rather than guesses.

Behavioral Measures and Analysis

Memory performance was quantified using two metrics: (1) *accuracy*, defined as the proportion of correct probe identifications, and (2) *bias scores*, calculated as the angular deviation between participants' memory reports and the original memory item, with higher scores indicating stronger distor-

Semantic	Visual	Mixed
aquatic activities/sea	balls/spherical	decorative/gold
baby toys/kid	rope/twine	green/vegetable
baked food/healthy	circular/discs	home tools/silver
body accessories/hair	groups of small objects/multicolored	ice/white
body part-related/brittleness	many colors/multicolored	natural minerals/crystal
containers for liquids/beverage	pattern/texture	paper-like/paper
cotton clothing/clothing	pole/sticks	plant/green
face accessories/eyes	red/shininess	pointy/tools
flammability/fire	string-like/netting	weaponry/silver
flying to not/flight	yellow/gold	wearable/black
ground animals/mammal		wood/brown
household furniture/furniture		
masculine/limbs		
medical supplies/absorbency		
music instruments/instruments		
natural resources/terrestrial		
old technology/used		
outdoor objects/outdoor		
recreational instruments/sport		
things with wheels/transportation		
wheeled vehicles/transportation		

Table 1: **Dimension Categorization.** Classify the 42-dimensional behavior-based object dimensions across semantic, visual, and mixed dimensions.

tion toward the induction direction. The bias score is calculated using the following formula:

$$\text{Bias Score} = |\theta_{\text{Report}} - \theta_{\text{Target}}| \tag{4}$$

where θ represents the angular coordinate. For example, if the angle deviation between the memory report and the original item is 15°, the bias score would be 15°.

Results

The experiment revealed systematic differences in memory vulnerability across dimensions (Figure 3b–c). For accuracy, visual dimensions (0.372) and mixed+visual combinations (0.484) showed significantly lower performance compared to semantic dimensions (0.503) and mixed+semantic combinations (0.498). Bias scores further highlighted this asymmetry: visual dimensions exhibited the strongest distortion (13.210°), followed by mixed+visual pairs (11.930°), whereas semantic dimensions (10.053°) and mixed+semantic pairs (9.194°) showed comparatively weaker biases.

Interpretation

These results demonstrate that perceptual comparisons with image similarity robustly induce retroactive memory distortions, specifically reflected in the tendency of participants to recall memory content that is biased toward the induced items with similar characteristics, with visual dimensions being disproportionately vulnerable. The gradient structure of the image wheels allowed us to isolate dimension-specific interference, confirming that even naturalistic, multidimensional objects are subject to similarity-driven biases. The weaker distortion in semantic dimensions aligns with theories positing that schematic or categorical representations stabilize memory traces against interference.

Dimension Wheels Induction Experiment

Stimuli and Task Design

Dimension wheels were constructed by embedding predefined activation values of visual or semantic dimensions into a latent space using an AI-driven generative model. Unlike image wheels, which interpolate between holistic perceptual

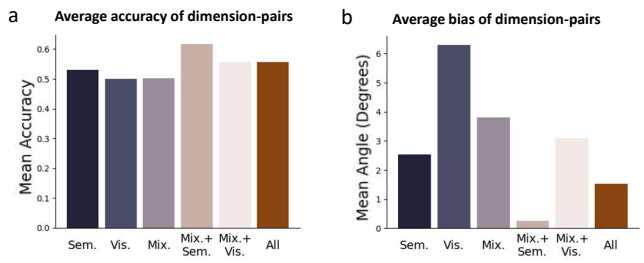

Figure 4: **Experimental results of the Dimension Wheels Induction Experiment(N=146).** (a) Mean accuracy of memory performance across dimension-pair categories. (b) Mean bias scores for dimension-pairs.

variations, dimension wheels explicitly manipulate isolated dimensions.

A total of 146 participants took part in this experiment on the Brain Island platform. The experimental procedure, total number of trials, and selection criteria for the dimension wheel experiment were the same as those for the image wheel experiment, with the only change being that all induction items were drawn from the dimension wheel.

Behavioral Measures and Analysis

As in the Image Wheels experiment, memory performance was assessed via *accuracy* (proportion of correct responses) and *bias scores* (angular deviation toward the induction item).

Results

The experiment revealed a pronounced hierarchy in dimension-specific memory vulnerability (Fig. 4). For accuracy, visual dimensions (mean accuracy = 0.500) and mixed+visual combinations (0.557) underperformed relative to semantic dimensions (0.530) and mixed+semantic pairs (0.617). Bias scores further emphasized this pattern: visual dimensions exhibited the strongest distortion (mean bias = 6.290°), followed by mixed pairs (3.805°) and mixed+visual pairs (3.090°), while semantic dimensions (2.530°) and mixed+semantic pairs (0.254°) showed minimal biases.

Interpretation

These results confirm that memory distortions can be induced even by abstract dimension activations, independent of holistic perceptual similarity. The significantly weaker biases in semantic dimensions compared to visual dimensions, suggest that semantic features benefit from conceptual hierarchies or schema-based stabilization.

In addition, although the overall memory distortion induced by the dimension wheel (accuarcy = 0.556, bias score = 1.520°) was smaller than that of the picture wheel (accuarcy = 0.470, bias score = 10.780°), the difference between the semantic dimension and the visual dimension was significantly larger in the dimension wheel. By decoupling dimension activations from perceptual context, this experiment advances computational models of VWM, highlighting the need to account for both feature-specific and integrative similarity mechanisms.

Discussion

Our findings provide novel insights into the dimensional architecture of vulnerability in visual working memory. By leveraging AI-driven generative models to disentangle and manipulate object dimensions in naturalistic stimuli, we addressed three critical gaps in the literature: (1) the construction of dimension-specific naturalistic memory probes, (2) the role of abstract dimension activations in inducing memory biases, and (3) the differential susceptibility of visual versus semantic features to retroactive interference.

Dimension-Specific Vulnerability in Memory. As predicted, both holistic (image wheels) and dimension-specific (dimension wheels) comparisons distorted memory. Visual dimensions showed greater distortion than semantic ones, supporting models that link semantic memory to deeper, schema-based processing (Brady et al., 2016; Chung et al., 2023). In contrast, early-stage visual features are more prone to interference from similar stimuli. This visual-semantic asymmetry challenges purely perceptual accounts of similarity-induced memory biases (Scotti et al., 2021; Saito, Kolisnyk, & Fukuda, 2023), suggesting that memory errors are also dimensionally organized. This has practical implications—for instance, enhancing semantic context could reduce memory distortions in settings like eyewitness testimony or AI-assisted tasks.

AI-Driven Methods for Cognitive Research. Our approach demonstrates how generative models can isolate and manipulate object dimensions in complex, realistic stimuli. Unlike traditional methods using simple shapes, our framework preserves naturalistic detail while enabling precise control. This allows for studying feature interactions in memory, and potentially in other areas like attention or decision-making. The distinction between image and dimension wheels further validates this method. While image wheels caused greater overall distortion, dimension wheels revealed a clearer divergence between visual and semantic memory errors. These findings point to the need for memory models to account for multidimensional similarity, not just perceptual distance.

Limitations and Future Directions. While our AI-driven approach advances stimulus control, several limitations warrant attention. First, our stimuli, though naturalistic, were constrained to predefined latent dimensions; real-world objects may exhibit emergent features not captured by current generative models. Second, our experiments focused on static comparisons, whereas dynamic interactions (e.g., sequential object manipulations) might reveal additional mechanisms of memory distortion. Finally, individual differences in cognitive style or expertise—factors known to influence schema formation—were not explored but could further explain variability in dimensional vulnerability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62472206), Shenzhen Science and Technology Innovation Committee (2022410129, KJZD20230923115221044, KCXFZ20201221173400001), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003), and the Center for Computational Science and Engineering at Southern University of Science and Technology.

References

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful. *Journal of cognitive neuroscience*, *33*(5), 902–918.

Bracci, S., & Op de Beeck, H. P. (2023). Understanding human object vision: a picture is worth a thousand representations. *Annual review of psychology*, *74*(1), 113–135.

Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 942.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (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.

Chung, Y. H., Brady, T. F., & Störmer, V. S. (2023). No fixed limit for storing simple visual features: Realistic objects provide an efficient scaffold for holding features in mind. *Psychological Science*, *34*(7), 784–793.

Chunharas, C., Rademaker, R. L., Brady, T. F., & Serences, J. T. (2022). An adaptive perspective on visual working memory distortions. *Journal of Experimental Psychology: General*, *151*(10), 2300.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in brain research*, *169*, 323–338.

Fukuda, K., Pereira, A. E., Saito, J. M., Tang, T. Y., Tsubomi, H., & Bae, G.-Y. (2022). Working memory content is distorted by its use in perceptual comparisons. *Psychological science*, *33*(5), 816–829.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Karras, T. (2019). A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science advances*, *9*(17), eadd2981.

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

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., & Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Muttenthaler, L., Zheng, C. Y., McClure, P., Vandermeulen, R. A., Hebart, M. N., & Pereira, F. (2022). Vice: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, *35*, 33661–33675.

Saito, J. M., Bae, G.-Y., & Fukuda, K. (2024). Judgments during perceptual comparisons predict distinct forms of memory updating. *Journal of Experimental Psychology: General*, *153*(1), 38.

Saito, J. M., Duncan, K., & Fukuda, K. (2023). Comparing visual memories to similar visual inputs risks lasting memory distortion. *Journal of Experimental Psychology: General*, *152*(8), 2318.

Saito, J. M., Kolisnyk, M., & Fukuda, K. (2023). Perceptual comparisons modulate memory biases induced by new visual inputs. *Psychonomic Bulletin & Review*, *30*(1), 291–302.

Sasin, E., Markov, Y., & Fougnie, D. (2023). Meaningful objects avoid attribute amnesia due to incidental long-term memories. *Scientific reports*, *13*(1), 14464.

Scotti, P. S., Hong, Y., Leber, A. B., & Golomb, J. D. (2021). Visual working memory items drift apart due to active, not passive, maintenance. *Journal of Experimental Psychology: General*, *150*(12), 2506.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751.

Wei, C., Zhang, C., Zou, J., Deng, H., Heinke, D., & Liu, Q. (2025). Synthesizing images on perceptual boundaries of anns for uncovering and manipulating human perceptual variability. *arXiv preprint arXiv:2505.03641*.

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024a). Cocog-2: Controllable generation of visual stimuli for understanding human concept representation. *arXiv preprint arXiv:2407.14949*.

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024b). Cocog: Controllable visual stimuli generation based on human concept representations. *arXiv preprint arXiv:2404.16482*.

Zheng, C. Y., Pereira, F., Baker, C. I., & Hebart, M. N. (2019). Revealing interpretable object representations from human behavior. *arXiv preprint arXiv:1901.02915*.