CoCoG: Controllable Visual Stimuli Generation Based on Human Concept Representations

Chen Wei[*,1,2] , **Jiachen Zou**[*,1] , **Dietmar Heinke**[2] and **Quanying Liu**[1]
[1]Southern University of Science and Technology, Shenzhen, China
[2]University of Birmingham, Birmingham, United Kingdom
{weic3, zoujc2022}@mail.sustech.edu.cn, d.g.heinke@bham.ac.uk, liuqy@sustech.edu.cn

Abstract

A central question for cognitive science is to understand how humans process visual objects, i.e, to uncover human low-dimensional concept representation space from high-dimensional visual stimuli. Generating visual stimuli with controlling concepts is the key. However, there are currently no generative models in AI to solve this problem. Here, we present the Concept based Controllable Generation (CoCoG) framework. CoCoG consists of two components, a simple yet efficient AI agent for extracting interpretable concept and predicting human decision-making in visual similarity judgment tasks, and a conditional generation model for generating visual stimuli given the concepts. We quantify the performance of CoCoG from two aspects, the human behavior prediction accuracy and the controllable generation ability. The experiments with CoCoG indicate that 1) the reliable concept embeddings in CoCoG allows to predict human behavior with 64.07% accuracy in the THINGS-similarity dataset; 2) CoCoG can generate diverse objects through the control of concepts; 3) CoCoG can manipulate human similarity judgment behavior by intervening key concepts. CoCoG offers visual objects with controlling concepts to advance our understanding of causality in human cognition. The code of CoCoG is available at https://github.com/ncclab-sustech/CoCoG.

1 Introduction

Humans receive abundant visual stimulation from the natural world. Unlike computer vision models that aim at object recognition tasks, humans aim to survive in complex nature, which requires understanding abstract concepts of these visual objects, such as functionality, toxicity, and danger. To explore such concept representation in humans, scientists have proposed a series of visual stimuli-based decision-making tasks [Murphy and Medin, 1985; Medin *et al.*, 1993; Hebart *et al.*, 2020; Roads and Love, 2023], such as the similarity judgment task, in which visual stimuli of specific concepts are presented to the participants and then their decision-making behaviors are recorded. However, understanding human concept representation through these cognitive tasks has two main difficulties. On the one hand, training an AI agent to uncover interpretable concept representations by predicting human decision-making behavior requires substantial human decision-making data under massive visual objects. On the other hand, to understand the causal relationship between concept representation and behavior, it is necessary to manipulate the concepts, preserving all other low-level features, to generate visual objects. This is unexplored territory for AI. It poses a new technical challenge for image generation, namely *controllable visual object generation based on concept representation*.

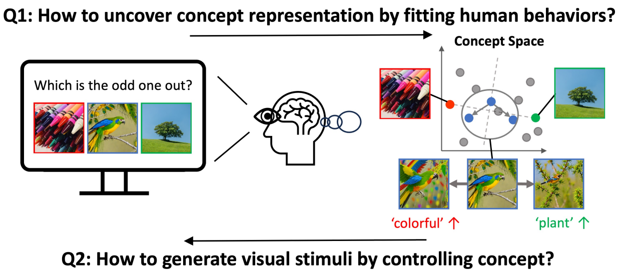
Q1: How to uncover concept representation by fitting human behaviors?
Concept Space
Which is the odd one out?
'colorful' ↑
'plant' ↑
Q2: How to generate visual stimuli by controlling concept?

Figure 1: Motivations of our work.

Controllable generation models have made great progress in AI community, for example, conditional generation models based on GAN [Goodfellow *et al.*, 2020; Tao *et al.*, 2022] and diffusion [Song *et al.*, 2020; Ho *et al.*, 2020; Ho and Salimans, 2022]. These conditional generation models have been applied in many fields, including text-to-image/video generation [Rombach *et al.*, 2022; Ramesh *et al.*,], image inverse problem [Kawar *et al.*, 2022; Chung *et al.*, 2022; Meng *et al.*, 2021] and biomedical imaging[Song *et al.*, 2021; Özbey *et al.*, 2023]. The conditions for controllable image generation can come from multiple aspects, such as text, sketches, edge maps, segmentation maps, depth maps [Rombach *et al.*, 2022; Ramesh *et al.*, ; Meng *et al.*, 2021; Zhang *et al.*, 2023; Yu *et al.*, 2023; Bansal *et al.*, 2023]. However, these conditions do not include human subjective feelings nor human feedback, resulting in generated images misaligned with human needs. For instance, generated im-

The appendix is available at https://arxiv.org/abs/2404.16482.

ages by recommendation system may not meet human preference. To align the generated images with human needs, some pioneer works brought human feedback (e.g., the human visual preference score obtained offline [Wu *et al.*, 2023; Kirstain *et al.*, 2023], the human-in-the-loop visual comparison decision [von Rütte *et al.*, 2023; Fan *et al.*, 2023; Tang *et al.*, 2023]) into conditional generative models. Nevertheless, these works have not consider prior knowledge of cognitive science, such as the factors that have the greatest impact on human decision-making, that is, concepts, as control variables for image generation.

A large number of human research suggest that similarity judgment tasks are an effective experimental paradigm for revealing human concept representations [Roads and Love, 2023; Hebart *et al.*, 2020]. In these tasks, human subjects are asked to compare different visual stimuli and make decisions based on their similarity. AI models, such as [Peterson *et al.*, 2018; Marjieh *et al.*, 2022b; Marjieh *et al.*, 2022a; Muttenthaler *et al.*, 2022a; Jha *et al.*, 2023; Fu *et al.*, 2023], are proposed to predict the subjects' decisions, under an assumption that human perception of visual objects can be encoded into a low-dimensional mental representation, namely *concept embedding*. The distance between visual objects in this concept space reflects the distance of visual objects in the human's mind. Human decision-making behavior can be explained by different dimensions of concept representation. Thus, aligning AI models with humans in terms of concept representation would naturally align their outputs as well. Also, compared with existing controllable image generation methods, a controllable generation model based on concept representation would control human decision-making behavior more effectively.

In this study, we propose a Concept based Controllable Generation (CoCoG) framework. CoCoG utilize concept embeddings as conditions for the image generation model, bridging cognitive science and AI. CoCoG comprises two parts: a *concept encoder* for learning concept embedding via predicting human behaviors and a *concept decoder* which employs a conditional diffusion model for mapping concept embedding to visual stimuli through a two-stage generation strategy.

CoCoG has three main contributions:

- CoCoG's concept encoder can predict human visual similarity decision-making behaviors with higher accuracy than the state-of-the-art (SOTA) model (i.e., VICE [Muttenthaler *et al.*, 2022b]), uncovering a reliable, interpretable concept space of humans (Figure 3).
- CoCoG's concept decoder can generate visual stimuli by controlling the concept embedding. The generated visual stimuli have diversity and high consistency with the target concept embeddings (Figure 4&5). The generation can be guided with text prompts (Figure 6&7).
- CoCoG can manipulate the human similarity decision-making behavior in a highly controllable way by controlling the concepts of generated visual stimuli (Figure 8).

2 Method

The CoCoG method comprises two parts: a concept encoder and a two-stage concept decoder, as shown in Figure 2.

2.1 Concept Encoder for Embedding Low-Dimensional Concepts

The first step in CoCcoG is to train a concept encoder to learn the concept embeddings of visual objects (Figure 2a). Given a dataset of visual objects X, for each visual object x in the dataset, we first input it into the CLIP image encoder f to extract the CLIP embedding $h \in \mathcal{R}^D$. Then, the CLIP embedding is input into a learnable concept projector g, thereby generating the concept embedding of the visual object $c \in \mathcal{R}^d$:

$$\begin{aligned} &\text{CLIP embedding:} && h = f(x), \\ &\text{concept embedding:} && c = g(h), \end{aligned} \tag{1}$$

where the each dimension in concept embedding c represents aninterpretable concept, and the corresponding activation of this dimension indicates the activation strength of the concept in the visual object. In short, we have $c = g(f(x))$.

To train this model, we used the *triplet odd-one-out* similarity judgment task in the THINGS dataset [Hebart *et al.*, 2023]. In this task, participants are asked to view three visual objects and consider the similarity between each pair of visual objects to determine the similar pair and select the remaining one as the *odd-one-out*.

Similar to previous works [Zheng *et al.*, 2018; Hebart *et al.*, 2020; Muttenthaler *et al.*, 2022b], we use the dot product similarity as the similarity measurement function between concept embeddings (i.e., $S_{ij} =< c_i, c_j >$) and use cross-entropy based on pairwise similarities to predict human decisions. Therefore, for the concept embeddings c_i, c_j, c_k of three visual objects x_i, x_j, x_k, we have

$$p(y) = CrossEntropy(S_{jk}, S_{ik}, S_{ij}), \tag{2}$$

where $p(y)$ is the probability distribution of the triplet visual stimuli (x_i, x_j, x_k) being the *odd-one-out*, with the highest probability choice being the model's predicted behavioral outcome. By comparing the model's predicted behavioral outcomes with the recorded behavioral outcomes from human participants, we calculate the loss and perform backpropagation to train the concept projector g. In the specific process, we added an L_1 regularization to c to ensure the sparseness of low-dimensional concept embedding. We verified different training strategies (see Appendix).

2.2 Two-Stage Concept Decoder for Controllable Visual Stimuli Generation

After training the concept encoder, we can obtain the data triplet (x_i, h_i, c_i) for each image in the dataset. Based on these data, we then train a concept decoder to generate visual objects by controlling human conceptual representations.

In our model, the concept embedding c completely determines the distribution of the human decision y, and the h completely determines the distribution of c. Therefore, the joint distribution of the human decisions $p(x, h, c, y)$ can be formulated as:

$$p(x, h, c, y) = p(y)p(c|y)p(h|c)p(x|h). \tag{3}$$

Next, we present each step of $p(c|y), p(h|c), p(x|h)$ respectively. $p(c|y)$ means finding the appropriate concept embedding c given the behavioral outcome y. This step is determined according to specific control objectives, which we will

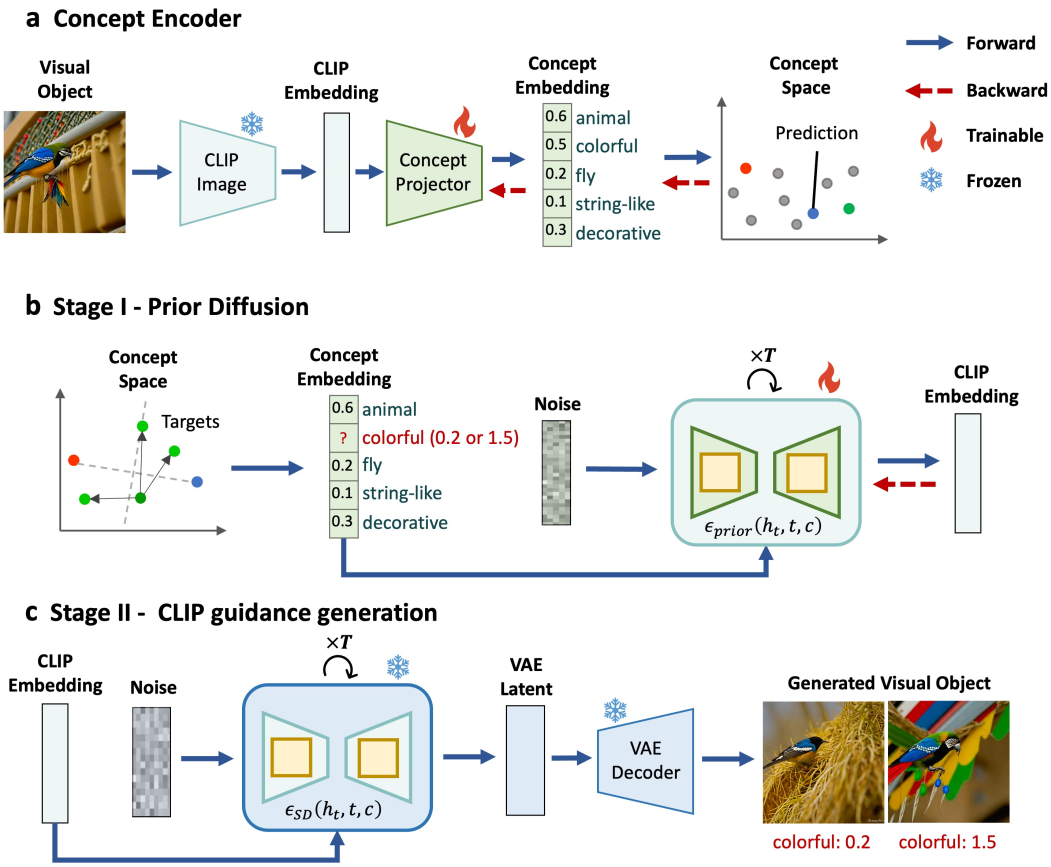

Figure 2: The framework of CoCoG. (a) The *concept encoder* for learning concept embeddings using a similarity judgment behavior dataset. Visual objects are processed through the CLIP image encoder to obtain CLIP image embeddings, and then passed through a learnable concept projector to obtain concept embeddings. Then, we can predict similarity judgment behaviors by compute similarity with others. (b) The stage I of the concept decoder, the *prior diffusion* for determining the concept embedding based on our desired judgment behavior (e.g., here modifying the concept "colorful"). Then, we train a diffusion model conditioned on the concept embedding to generate the corresponding CLIP embedding. (c) Stage II of the concept decoder, the *CLIP guided generation*. It uses the CLIP embedding as a condition to guide the pre-trained image diffusion generation model to generate VAE latent, which are then processed through the VAE decoder to produce the generated visual object.

discuss in chapter 4.2. For $p(c|y)$ and $p(x|h)$, we decompose the concept decoder into two stages, the Prior Diffusion and the CLIP guided generation, respectively, which execute the processes of generating CLIP embedding h from the concept embedding c and generating visual object x from the CLIP embedding h.

Stage I - Prior diffusion. Inspired by DALL-E 2 [Ramesh *et al.*,], we train a diffusion model conditioned on the concept embedding c to learn the distribution of CLIP embeddings $p(h|c)$ (Figure 2b). The CLIP embedding h obtained in this stage serves as the prior for the next stage. We construct a lightweight U-Net: $\epsilon_{prior}(h_t, t, c)$, where h_t represents the noisy CLIP embedding in the diffusion time step t. We extract the CLIP embeddings and the concept embeddings from ImageNet and use the obtained training pairs (h_i, c_i) to train the Prior Diffusion model. We employ the classifier-free guidance method to train this conditional generative diffusion model. The specific formulas can be found in the Appendix.

Stage II - CLIP guidance generation. After obtaining the CLIP embedding h in Stage I, we model a generator $p(x|h)$ to sample the visual object x conditional to h (Figure 2c). In this study, we use the pre-trained SDXL and IP-Adapter models[Podell *et al.*, 2023; Sauer *et al.*, 2023; Ye *et al.*, 2023]. Speicfically, SDXL serves as the backbone of the image diffusion generation model. Through the dual cross-attention modules of the IP-Adapter, we input the CLIP embedding h as a condition, thereby guiding the denoising process of the U-Net. Similar to Stage I, we have the model $\epsilon_{SD}(z_t, t, h)$, where z_t are the noisy latents of SDXL's VAE. Details can be found in the Appendix. Since we freeze the pre-trained models without any modification, we can simply use the existing functionalities of the pre-trained models and combine them with concept embedding guidance.

CLIP embedding as an intermediate variable. In both processes from visual objects to concept embeddings and from concept embeddings back to visual objects, we use CLIP embeddings as an intermediate variable. This is for

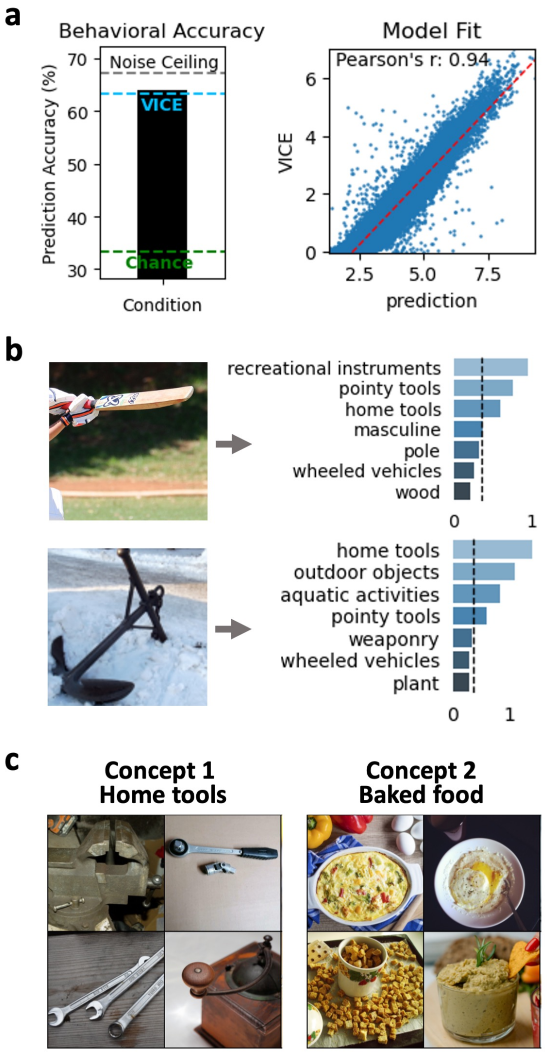

Figure 3: The performance of the concept encoder in predicting and explaining human behavior. (a) Our model's prediction accuracy for similarity judgment behavior is 64.07%, exceeding the previous SOTA model VICE's 63.27% (blue dashed line), with only slightly lower than the noise ceiling (gray dashed line) [Hebart *et al.*, 2020]. The Pearson correlation coefficient between the similarity of visual objects predicted by our model and by VICE is 0.94; (b) Example visual objects and their concept embeddings, with dashed lines representing the 90th percentile of activated concepts; (c) Example visual objects with significant activation on the concept *Home tools* and *Baked food*, respectively.

two purposes: 1) For the concept encoder, CLIP embeddings are sufficiently low-dimensional and contain key information of images. Previous study has shown that using CLIP embeddings along with simple linear probing can well predict human behavior in similarity judgment tasks [Muttenthaler *et al.*, 2022a]. 2) For the concept decoder, using CLIP embeddings helps us better utilize existing pre-trained conditional generative models. Existing models can already use CLIP embeddings as a conditional input for the generative model, allowing us to simply adopt a two-stage generation strategy. We only need to train a Prior Diffusion model, which significantly reduces the computational cost of training and inference.

3 Model Validation

3.1 Concept Encoder Can Predict and Explain Human Behaviors

We first validated the concept encoder from two aspects: the prediction accuracy of human behaviors (Figure 3a) and the interpretability of the learned lwo-dimensional concepts (Figure 3b&c). We used the THINGS Odd-one-out dataset as the similarity judgment behavior dataset to train our concept encoding model. We tested various configurations of the concept encoding model (which will be detailed in the Appendix). Figure 3a shows the results of the optimal model configuration. We used 42-dimensional concept embeddings for comparison with previous SOTA model (VICE) and our experiments have found that more than 42 dimensions only bring small improvements. In terms of behavioral prediction, our best-performing model achieved an accuracy of 64.07% on the THINGS Odd-one-out dataset, surpassing the previous best model VICE's accuracy of 63.27%. We also compared the similarity predictions of visual objects between our model and VICE, and the Pearson correlation coefficient of their predictions reached 0.94. This indicates that our model can accurately predict human similarity judgment behaviors.

One major merit of our concept embedding is its good interpretability. Figure 3b shows the low-dimensional concept embeddings for the two example visual objects. It is obvious that the concept embeddings encoded by CoCoG well describe the visual objects, and the activation of concepts in a visual object exhibits a favorable property of sparsity, which is in line with human intuition (see Appendix). We used the CLIP-Dissect model to automatically choose words from the lexicon to describe the concepts in each of the 42 dimensions. Note that the relationship between these words and the concepts is not absolute and is only for examples. Figure 3 shows the visual objects with the highest activation in the first two conceptual dimensions (i.e., home tools and baked food). Importantly, we find that the visual objects that significantly activate these dimensions are highly consistent with these two concepts. This indicates that our model can effectively encode the conceptual embeddings of visual objects, and the encoded concepts have good interpretability.

3.2 Concept Decoder Can Generate Visual Objects Consistent With Concept Embedding

In this section, we validate the generative effectiveness of the concept decoder. Specific training parameters are shown in the Appendix. Figure 4 shows visual objects generated under the guidance of concept embeddings. These visual objects generated from the same concept embedding are well-aligned with the concept embedding and have good diversity, demonstrating that our diffusion model can conditionally generate visual objects consistent with the concept embeddings.

We quantify the similarity between the concept embeddings of the generated visual objects and the target concept

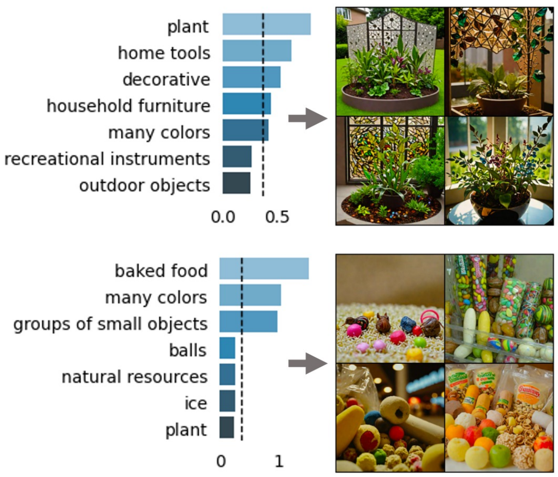

Figure 4: The visual objects generated by controlling the concept embeddings.

embeddings according to Eq.2. The similarity between generated visual objects and target concept embeddings is significantly higher than that between two random visual objects (Figure 5a). Additionally, by adjusting the guidance scale, we can control the guiding strength of the target concept embeddings, thereby controlling the similarity and diversity of the generated visual objects (specific metrics calculation can be found in the Appendix). As the guidance scale increases, the similarity between the visual objects and the target concept embeddings increases while the diversity decreases (Figure 5b).

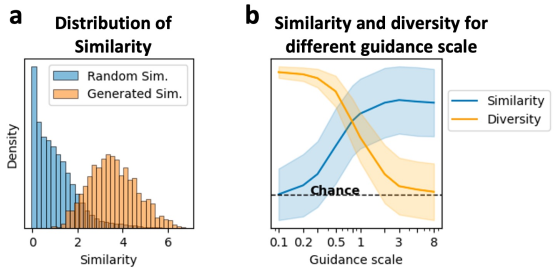

Figure 5: Measurements of the performance of generated visual objects. (a) The similarity between random visual objects and the target concept embeddings (blue), and the similarity between generated visual objects and the target concept embeddings (orange); (b) The similarity between visual objects and target concept embeddings as the guidance scale changes (blue), and the diversity of visual objects as the guidance scale changes (orange).

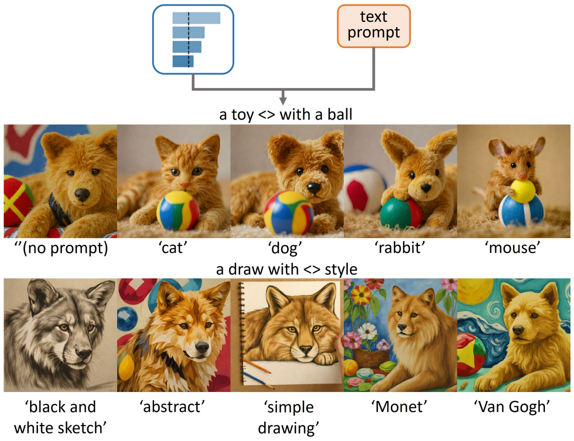

Figure 6: Visual objects generated with the same concept embedding combined with different text prompts. The first image in the first row is generated without using any text prompt, while other images are generated by using different animal species as text prompts. The images in the second row are generated by using different artistic styles as text prompts.

4 CoCoG for Studying Counterfactual Explanations of Human Behaviors

According to the role of concept embedding, manipulating the concept embedding can directly influence human similarity judgment behavior. Conversely, does the change in visual objects that do not affect the concept embedding have no impact on human behavior? CoCoG is an excellent tool to explore this counterfactual question.

4.1 Flexible Controlling of Generated Objects With Text Prompts

As shown in 6, we used the same concept embedding combined with different text prompts to generate visual objects. Regardless of the changes in the category and style of the visual objects, the images consistently retained the characteristics of the concept embedding. Note that we fixed the random seed to highlight the differences in text prompts. In theory, these completely different visual objects would lead to similar judgment behaviors in similarity judgment experiments.

Further, we show visual objects generated using the same text prompt combined with different concept embeddings (Figure 7). The images in the upper half were used to extract concept embeddings, which were then combined with text prompts to generate the new images in the lower half. The generated images reflect the property ”teddy bear” after adding the text prompt, with preserving the concept embeddings. Therefore, contrary to the concept-editing experiment, modifying text prompts do not lead to changes in judgment behavior because the concept embedding is preserved.

4.2 Manipulating the Similarity Judgment Decisions by Intervening the Key Concepts

In this experiment, we designed a simple scenario to show how CoCoG can manipulate similarity judgment behavior. Suppose participants need to perform a *Two alternative forced choice* experiment, i.e., they need to choose from Reference 1 & 2 which is more similar to the Query. In this scenario, we can directly increase or decrease certain concepts to make the Query more similar to one of the references. Considering that we use dot product similarity, we can simply choose the concept that is significant in the Reference but not

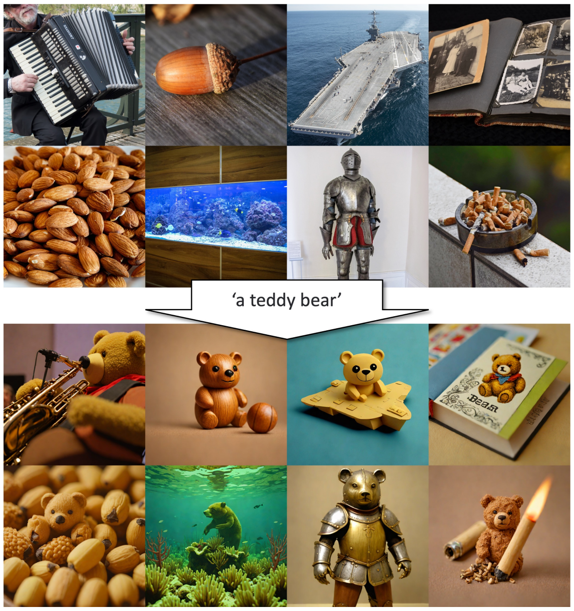

Figure 7: Visual objects generated with different concept embeddings combined with the same text prompt. The images in the upper and lower halves correspond to each other. The images in the upper half were used to extract concept embeddings, which were then combined with the text prompt 'a teddy bear' to generate the new images in the lower half.

in the Query as the key concept. The results in Figure 8, show that the preference of the Query for Reference 1 & 2 is manipulated by modifying two and three concepts, with the concepts shifting to two different directions. It is evident that the generated visual objects by CoCoG can change smoothly but significantly with the manipulation of concepts.

This method provides us with an effective tool for analyzing the causal mechanisms of concepts in similarity judgment tasks. For example, in the experiment, we can directly manipulate the concepts of interest and observe subsequent behavior. Researchers can actively choose to explore the impact of certain concepts on behavior and analyze the causal relationship between changes in concepts and judgment behavior. By focusing on informative concepts and precisely controlling the activation values of concepts, this method is expected to improve the efficiency of data collection and significantly reduce the number of experimental trials needed (rather than using millions of trials [Hebart *et al.*, 2023]).

5 Related Works

5.1 Concept Embeddings

Numerous computational models have been developed to model embeddings encoding in similarity judgment tasks [Roads and Love, 2023; Hebart *et al.*, 2020]. These methods aim to study concept embeddings through similarity judgment tasks, either by optimizing concept embeddings for each object [Zheng *et al.*, 2018; Roads and Love, 2021; Hebart *et al.*, 2020; Muttenthaler *et al.*, 2022b] or by using DNNs to predict human similarity judgment behavior [Peterson *et al.*, 2018; Marjieh *et al.*, 2022b; Marjieh *et al.*, 2022a; Muttenthaler *et al.*, 2022a; Jha *et al.*, 2023; Fu *et al.*, 2023]. Previous methods, assuming sparsity, continuity, and positivity of concept embeddings, learn low-dimensional concept embeddings for each object through probabilistic models. These embeddings have good interpretability but cannot generalize to new objects. On the other hand, DNN-based methods construct concept embeddings from neural activations by reducing the dimensionality of high-dimensional DNN latent representations [Jha *et al.*, 2023], aligning DNN-based similarity judgments with human judgments [Muttenthaler *et al.*, 2023], or using multimodal inputs to improve similarity judgment predictions [Marjieh *et al.*, 2022b; Marjieh *et al.*, 2022a]. This approach can utilize state-of-the-art DNN models to form embeddings that naturally generalize to new objects. To enhance behavioral experiments, Roads et al. utilized active learning coupled with a trial selection strategy for efficient concept embedding inference [Roads and Love, 2021]. Similarly, DreamSim leveraging Stable Diffusion, generates synthetic data within specified categories to create a Perceptual dataset, aimed at studying Human perceptual judgments [Fu *et al.*, 2023].

5.2 Conditional Diffusion Models

Recently, conditional generative models based on diffusion models have seen significant development. The classifier-free guidance method has demonstrated strong controllable generative capabilities through supervised learning [Ho and Salimans, 2022]. These methods dominate the tasks of text-to-image generation. From the widely used Stable Diffusion [Rombach *et al.*, 2022; Podell *et al.*, 2023; Sauer *et al.*, 2023] to subsequent methods like ControlNet [Zhang *et al.*, 2023] and IP-Adapter [Ye *et al.*, 2023], they have added more controllable conditions to conditional generation, such as edge maps, segmentation maps, depth maps [Rombach *et al.*, 2022; Ramesh *et al.*, ; Meng *et al.*, 2021; Zhang *et al.*, 2023; Yu *et al.*, 2023; Bansal *et al.*, 2023]. Additionally, conditional generative models based on human feedback and preferences have also shown potential. Works like FABRIC and DPOK use real-time similarity tasks to guide models to generate images meeting human needs, proving that human feedback can provide more nuanced control over conditional generative models [von Rütte *et al.*, 2023; Fan *et al.*, 2023; Tang *et al.*, 2023]. Methods like Pick-a-pic and Human Preference Score generate images aligned with human aesthetics [Kirstain *et al.*, 2023; Wu *et al.*, 2023]. These advancements demonstrate that, in comparison to standard text-to-image frameworks, conditional generative models possess significant potential for more closely meeting human needs and preferences.

6 Discussion

We proposed the CoCoG model, capable of predicting human visual similarity judgment behavior and learning human conceptual embeddings for visual objects. It can also efficiently and controllably generate visual objects in line with human cognition (Fig 3), manipulating human similarity judgment behavior and studying causal mechanisms in human cognition (Fig 8).

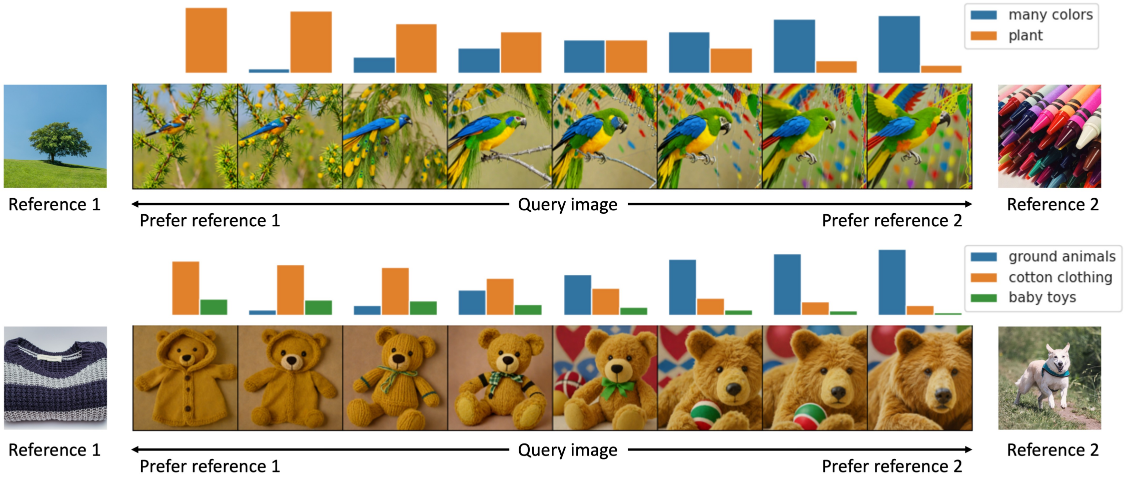

Figure 8: Manipulating similarity judgment behavior by key concepts intervention. (upper) The visual objects generated by using "many colors" and "plant" as key concepts. (bottom) The visual objects generated by using "ground animals", "cotton clothing", and "baby toys" as key concepts.

Contributions to AI. Our approach bridges generative models and human visual cognition. Through the concept encoding model, we align DNNs with human visual concept representations, simulating human processing and responses to visual objects more precisely, with potential to enhance AI's visual understanding capabilities; through controllable diffusion generation based on concept embeddings, we make conditional generative models more closely linked to human cognition and can manipulate human behavior through generated stimuli, promising to improve control and safety in AI-human interactions.

Contributions to cognitive science. Our approach significantly expands research on human visual cognition in cognitive science. With the concept encoding model, we achieved interpretable concept encoding for visual objects (Figure 3b); with controllable diffusion generation based on concept embeddings, we can generate a rich variety of natural stimuli to control human similarity judgment behaviors (Figure [Roads and Love, 2021; Fu *et al.*, 2023]). By combining advanced AI models with cognitive science research methods, we greatly enhance the efficiency and breadth of visual cognition research. Additionally, this method may reveal causal mechanisms in human visual cognition, offering new perspectives for understanding human cognitive processes.

Future directions. In the future, we will extend the paradigm of human visual cognition research to the study of AI representational spaces, which would help align AI with humans and provide new insights for understanding AI's cognition. Also, we recommend to bring optimal experimental design [Rainforth *et al.*, 2023; Roads and Love, 2021] into human experiments. It will largely improve the efficiency of human behavioral data collection, facilitate model learning, optimizing experimental paradigms with broader applications in cognitive science and AI.

Ethical Statement

The human behavioral data in this study is from THINGS public dataset. No animal or human experiments are involved.

Acknowledgments

This work is supported by the National Key R&D Program of China (2021YFF1200804), Shenzhen Science and Technology Innovation Committee (20200925155957004, KCXFZ20201221173400001, SGDX2020110309280100), Guangdong Provincial Climbing Plan (pdjh2024c21310).

Contribution Statement

Chen Wei and Jiachen Zou contribute equally to this work. Quanying Liu is the corresponding author.

References

[Bansal *et al.*, 2023] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[Chung *et al.*, 2022] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

[Fan *et al.*, 2023] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023.

[Fu *et al.*, 2023] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Hebart *et al.*, 2020] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

[Hebart *et al.*, 2023] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jha *et al.*, 2023] Aditi Jha, Joshua C Peterson, and Thomas L Griffiths. Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive science*, 47(1):e13226, 2023.

[Kawar *et al.*, 2022] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[Kirstain *et al.*, 2023] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

[Marjieh *et al.*, 2022a] Raja Marjieh, Ilia Sucholutsky, Theodore R Sumers, Nori Jacoby, and Thomas L Griffiths. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*, 2022.

[Marjieh *et al.*, 2022b] Raja Marjieh, Pol Van Rijn, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. Words are all you need? language as an approximation for human similarity judgments. In *The Eleventh International Conference on Learning Representations*, 2022.

[Medin *et al.*, 1993] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993.

[Meng *et al.*, 2021] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Murphy and Medin, 1985] Gregory L Murphy and Douglas L Medin. The role of theories in conceptual coherence. *Psychological review*, 92(3):289, 1985.

[Muttenthaler *et al.*, 2022a] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2022.

[Muttenthaler *et al.*, 2022b] Lukas Muttenthaler, Charles Y Zheng, Patrick McClure, Robert A Vandermeulen, Martin N Hebart, and Francisco Pereira. Vice: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, 35:33661–33675, 2022.

[Muttenthaler *et al.*, 2023] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew K Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. *arXiv preprint arXiv:2306.04507*, 2023.

[Özbey *et al.*, 2023] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

[Peterson *et al.*, 2018] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[Rainforth *et al.*, 2023] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *arXiv preprint arXiv:2302.14545*, 2023.

[Ramesh *et al.*,] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents.

[Roads and Love, 2021] Brett D Roads and Bradley C Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3547–3557, 2021.

[Roads and Love, 2023] Brett D Roads and Bradley C Love. Modeling similarity and psychological space. *Annual Review of Psychology*, 75, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn

Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Sauer *et al.*, 2023] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[Song *et al.*, 2021] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2021.

[Tang *et al.*, 2023] Zhiwei Tang, Dmitry Rybin, and Tsung-Hui Chang. Zeroth-order optimization meets human feedback: Provable learning via ranking oracles. *arXiv preprint arXiv:2303.03751*, 2023.

[Tao *et al.*, 2022] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.

[von Rütte *et al.*, 2023] Dimitri von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. Fabric: Personalizing diffusion models with iterative feedback. *arXiv preprint arXiv:2307.10159*, 2023.

[Wu *et al.*, 2023] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[Ye *et al.*, 2023] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[Yu *et al.*, 2023] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *arXiv preprint arXiv:2303.09833*, 2023.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zheng *et al.*, 2018] Charles Y Zheng, Francisco Pereira, Chris I Baker, and Martin N Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, 2018.