Semantic bias in image-text matching in humans versus vision-language pretraining AI models

Jinhan Zhang[1], Qichun Duan[2], Chenyang Zhao[2], Antoni B. Chan[4], Janet H. Hsiao[2,3]
[1] Division of Emerging Interdisciplinary Areas, HKUST
[2] Division of Social Science, HKUST
[3] Department of Computer Science & Engineering, HKUST
[4] Department of Computer Science, City University of Hong Kong
{jinhan.zhang, qduanac}@connect.ust.hk, {zhaocy, jhhsiao}@ust.hk, abchan@cityu.edu.hk

Abstract

Recent research has shown that vision-language pretraining (VLP) models using contrastive learning (Contrastive Language Image Pretraining, CLIP) has a semantic bias towards using concrete words during image-text matching. Here we showed that as compared with CLIP, humans attended more to abstract words during image-text matching. This difference likely results from CLIP's difficulty in developing grounded understanding of abstract concepts and capturing contextual dependencies between images and captions through contrastive learning. While CLIP's caption attention aligned more closely with humans for concrete than for abstract captions, their alignment in image attention did not differ between the caption conditions, as human image attention was driven primarily by individual differences in explorative or focused perceptual style. Our findings thus revealed important differences in information processing mechanisms between humans and the VLP models, with important implications for not only human and AI image-text matching, but also for potential ethical issues resulting from such misalignment.

Keywords: image-text matching; word concreteness; vision-language model; explainable AI; human attention; eye tracking

Introduction

The association and alignment between visual and linguistic information are fundamental to human cognition, supporting robust understanding of the world and enabling key abilities such as vocabulary acquisition and concept learning. Recently, to equip AI with multimodal learning capabilities, Vision–Language Pretraining (VLP) models have been proposed to learn visual-linguistic associations. Among these, Contrastive Language–Image Pretraining (CLIP) is one of the most prominent models. CLIP adopts a dual-encoder architecture and leverages contrastive learning for image–text matching (i.e., aligning matched image–text pairs while separating mismatched pairs in a shared embedding space) on web-scale image-text datasets, resulting in robust representations. As a foundation model, CLIP has been widely applied to downstream tasks such as image–text retrieval and visual question answering in zero-shot and fine-tuning settings (Changpinyo et al., 2021; Luo et al., 2022).

Despite CLIP's impressive performance, the complexity of its deep-learning and transformer architecture has made it difficult to understand the internal decision-making mechanisms. Recent advances in explainable AI (XAI) have proposed methods to visualize features contributing to model outputs. Class-activation map–based approaches such as Grad-CAM (Selvaraju et al., 2017) have been widely used for interpreting convolutional neural networks, but they often yield low-faithfulness explanations when applied to transformer architectures due to architectural differences (Zhao et al., 2025). Alternative transformer-specific methods have been proposed. However, these methods typically treat VLP as a vision transformer without capturing the contrastive image–text matching objective (e.g., Attention Rollout, Abnar & Zuidema, 2020) and generate explanations based on self-attention (e.g., GAME, Chefer et al., 2021), which provides a bottom-up explanation that is agnostic to the task, often leading to confusing saliency maps due to the sparse attention maps. Similarity-based approaches (e.g., MaskCLIP, Zhou et al., 2022) that compute cosine similarity between image and text features address some of these limitations but remain limited in explaining the bottom-up processes and primarily reflect only image-level importance. To overcome these issues, Zhao et al. (2024) proposed Grad-ECLIP, a gradient-based method that identifies image regions and words that most strongly influence the model's matching output and produces more faithful explanations for both the image and text pathways than previous methods.

Using Grad-ECLIP to analyze CLIP's information use in image–text matching, Zhao et al. (2025) identified a semantic bias toward concrete words, evidenced by a positive correlation between word concreteness and their importance scores in contributing to CLIP's output. This finding suggests that CLIP may exhibit weaker representations of abstract words and concepts, potentially leading to greater difficulty in processing abstract captions and matching them to images.

CLIP's weaker representations of abstract compared with concrete concepts may arise from the nature of the image–text association task. Concepts with high concreteness and imageability—such as colors and object shapes—are easier to identify in images, allowing the model to learn concept–feature associations from regularities in the training data. This pattern parallels human word processing, where concrete words are learned more easily due to their grounding in sensorimotor experiences in the physical world (Paivio et al., 1986). By contrast, learning abstract concepts depends more on linguistic information and word associations (Borghi et al., 2011), which are not directly observable from images depicting similar concepts. Accordingly, abstract words are generally more difficult to process than concrete words, often requiring longer processing times as reported in many prior works (Bleasdale, 1987; Schwanenflugel et al., 1989). However, it remains unclear whether such processing

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

difficulty influences adults' attention patterns in the image–text matching task and induces a concreteness bias in humans' text processing, similar to that observed in CLIP.

Another factor that may influence humans' attention to images during the matching task is individual visual scanning preferences. Past research has documented a consistent individual difference in how people extract information from visual scenes, with two characteristic attentional strategies identified: an explorative strategy, involving scanning broader areas that may involve both the foreground and background, and a focused strategy, involving narrower, more concentrated scanning with a focus on foreground objects or features. These patterns have been observed across various image-viewing tasks, including passive viewing (Hsiao et al., 2021), object classification (Qi et al., 2024), and object detection (Yang et al., 2023). It remains unclear, however, whether these representative strategies persist in the image–text matching context, and if so, whether strategy choice is influenced by differences in task demands when viewing abstract versus concrete image–text pairs.

Comparing attention patterns between AI and humans is essential for developing more interpretable and trustworthy models. Recent research suggests that, due to modern AI's emergent intelligent behavior and humans' theory-of-mind abilities in social interactions, people interacting with AI typically initially assume that it processes information similarly to themselves, and then update this belief after observing discrepancies (Yang et al., 2021; 2022; Hsiao et al., 2023). This highlights the importance of understanding processing differences between AI and humans, which are often obscured by the black-box nature of deep-learning models, in order to guide better human–AI alignment and mitigate potential ethical consequences (Hsiao, 2026).

In this study, we systematically examined differences in information use between AI and humans during image–text matching. We first examined whether the semantic bias toward concrete words observed in CLIP is also present in humans during image–caption matching and compared the strength of this effect between humans and CLIP. Given CLIP's attention bias toward concrete content and its potential difficulty in processing abstract language, we next assessed the alignment of CLIP's attention to the image and caption with human attention, asking whether CLIP's attention more closely resembles humans' when processing image–caption pairs with concrete captions than with abstract captions. To measure the information use of these two agents, we collected human attention data via eye-tracking and generated saliency maps highlighting the image regions and words important for CLIP's matching decisions using the XAI method Grad-ECLIP. Finally, building on prior works identifying two characteristic human visual attention styles—explorative and focused—when viewing images across various tasks, we tested if these individual differences persist in the current image–text matching context, if the attention style is associated with the different demands when viewing abstract vs concrete image-caption pairs, and if CLIP's visual attention aligns more closely with one style than the other.

Methods

Experiment Design

Participants We recruited 60 participants (28 female, 32 male) aged from 19 to 39 (M=22.07, SD=3.73) with normal or corrected-to-normal vision. All participants were proficient in English as verified by their LexTALE score with a mean score of 79.59% (Lemhöfer & Broersma, 2012).

Materials We sampled 100 images from the MSCOCO validation set (Lin et al., 2014), each containing 5 human-annotated captions, and selected two captions per image: one concrete and one abstract. From the total set of 100 images, a subset of 20 was paired with mismatched captions to encourage active engagement in the matching task, and data from these trials were excluded from all primary analysis.

To select captions, we first computed sentence concreteness for each caption as the mean concreteness of its words, using a human-rated word concreteness database (Brysbaert et al., 2014). For each image, the caption with the highest concreteness was chosen as the concrete caption and the one with the lowest as the abstract caption. To ensure that any observed effect could be attributed to concreteness rather than any other factors, a range of stimulus attributes was controlled, including caption length, word frequency in the training set, word importance, saliency map variance by Grad-ECLIP, and image-caption matching score predicted by CLIP, to match the two conditions.

Importantly, controlling CLIP's predicted matching score is necessary when examining alignment with human attention. Our preliminary analysis found a small but significant positive correlation between CLIP's matching score and sentence concreteness ($R^2 = 0.02$, $p < .001$). Without controlling for matching score, lower alignment with humans for abstract captions could simply reflect CLIP treating these pairs as less well matched, rather than a true difference in attention to abstract versus concrete words.

Apparatus The experiment was programmed using SR Experiment Builder (version 2.6.11) and run on a PC with Windows 10. Stimuli were displayed on a 15.6-inch FHD Monitor (1280 x 1024 resolution). Each image was resized to fit within a 735 x 736 pixel square frame centered on the screen, with space above and below. The caption was placed beneath the frame and, following Underwood et al. (2004), presented at 5 characters per visual degree at a viewing distance of 69 cm. Participants' eye movements were recorded by an EyeLink 1000 Plus eye tracker, installed on a tower mount and set to remote mode with a chin rest to minimize head movement. A standardized 9-point calibration procedure was performed before the experiment and repeated whenever the drift check exceeded 1° of the visual angle.

Procedure In the experiment, participants were instructed to view one image-caption pair at a time and rate the extent to which they considered the pair semantically matched. The

200 trials were organized as 4 blocks, each with 40 matched trials (20 concrete) and 10 mismatched trials. The concrete and abstract versions of the same image were always assigned to different blocks. Stimuli were presented in a randomized order both at the block level and within each block. A practice block consisting of 3 trials was run before the formal experiment to ensure that they understood the task. Each trial began with a drift check at the screen center, followed by a fixation cross at the center of the screen/image area for 500 ms. The image and the caption were then displayed accordingly. After finishing viewing the pair, participants pressed a key to proceed to a rating page, on which they used the mouse to drag a slider and rate how well the caption described the image on a scale from 0 (mismatched) to 100 (perfectly matched) (Figure 1). Participants' eye movements recorded from the onset of stimulus display until the key press were used for analysis.

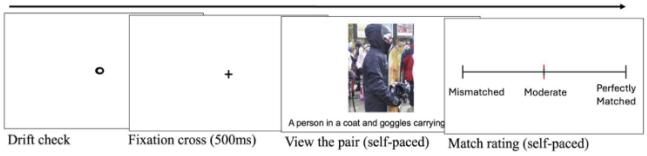

Figure 1: Experiment procedures

Attention Behavior Measures

CLIP Attention Maps We applied Grad-ECLIP to generate saliency maps that capture the contribution of image regions and words of the text caption to the model's prediction (Zhao et al., 2024; 2025). Grad-ECLIP is a gradient-based XAI method that explains image–text matching decisions in VLPs by attributing importance to pixels and tokens with respect to the predicted matching score. Unlike prior methods that rely on CLIP's sparse self-attention maps, Grad-ECLIP integrates channel-wise gradient importance with spatial (image-region and token-level) information, yielding dense, result-relevant importance maps that have been shown to outperform existing approaches. Following Zhao et al. (2025), we averaged subword tokens to generate word-level importance and applied min–max normalization to word importances within each sentence (See Figure 2 left for examples).

Human Attention Maps To compare with CLIP's visual saliency maps, we generated human attention heatmaps on the image by convolving each fixation point with a Gaussian kernel (SD = 41 pixels, corresponding to 1° of visual angle). To compare with CLIP's word importance values, we calculated the number of fixations on each word and applied min-max normalization (See Figure 2 right for examples).

Individual Difference in Eye Movement Patterns We used Eye Movement analysis with Hidden Markov Models (EMHMM) with co-clustering (Hsiao, Lan et al., 2021) to quantify individual differences in eye movement pattern when viewing the images across conditions. Taking both the spatial (fixation locations) and temporal (fixation transitions) information into account, each participant's eye movements on each stimulus were summarized by an hidden Markov model (HMM), which contains the person-specific regions of interest (ROIs) on that stimulus as the hidden states and transition probabilities among the ROIs. The optimal number of ROIs was determined from a range of 1 to 10 using a variational Bayesian approach. Co-clustering was then performed to group participants with similar visual scanning patterns to one another across stimuli. A representative HMM was generated to summarize eye movement patterns of participants in each group for each stimulus, and the number of ROIs in the representative HMM was set to be the median of group members' numbers of ROIs. The optimal number of participant groups was determined from 1 and 2 using a variational Bayesian approach. Both the individual HMMs and group HMMs during co-clustering were trained 500 times, and ones with the highest log-likelihood were chosen.

Following previous studies (e.g., Hsiao et al., 2021; Qi et al., 2024), we quantified each participant's eye movement patterns using the A–B scale, defined as $(L_A - L_B) / (|L_A + L_B|)$. Here, L_A and L_B denote the log likelihoods of a participant's eye movement data being clustered into Pattern Group A and B respectively, averaged over stimuli. Higher A–B values indicated greater similarity to Pattern Group A.

Data Analysis

Concreteness Bias Measurement We examined whether the previously reported semantic bias toward concrete words in CLIP was present in the current sampled dataset and tested whether the same pattern existed in humans. We first fit a linear regression model for each agent (CLIP and human) separately to test if word concreteness predicted the attention allocated. The same word in different sentences is treated as different samples. To compare the effect between CLIP and humans and examine whether they differed in attention bias towards concrete words, we constructed a linear mixed-effects model with word importance as the dependent variable, word concreteness, agent (CLIP vs. human, with human as the baseline), and their interaction as fixed effects, and image as the random effect.

CLIP-Human Attention Alignment We measured attention pattern similarity between CLIP and humans using the Pearson correlation coefficient (PCC), computed separately for flatten image attention and caption attention. Attention data across all human participants were aggregated for each image or caption to compute the attention alignment with CLIP. To test our hypothesis of whether human and CLIP attention patterns were more closely aligned when matching image-caption pairs in the concrete condition than in the abstract condition, we fitted a linear mixed effects model with CLIP–human attention alignment as the dependent variable, condition (concrete vs. abstract, with abstract as the baseline) as a fixed effect, and image as a random effect. We constructed the same model for visual attention and textual attention as the dependent variable separately.

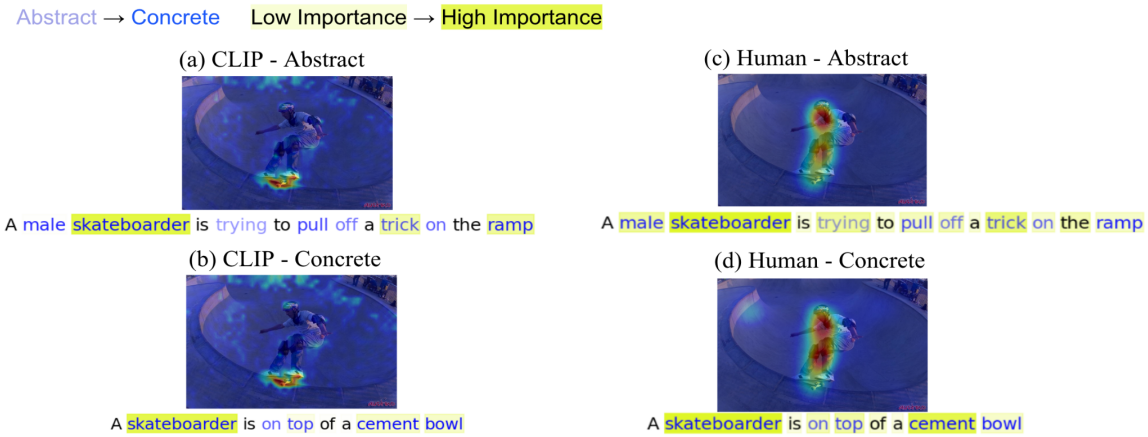

Figure 2: Examples of CLIP's (left) and humans' (right) image attention (heatmaps) and caption attention (words with importance highlight) when matching images with abstract captions (top) vs concrete captions (bottom). Words with a darker yellow background indicate higher importance for the matching decision. Words printed in a darker blue shade indicate higher concreteness. Removed stop words and words not found in the concreteness database are shown in black.

To examine whether CLIP had better attention alignment with human participants of a particular perception style (Hsiao et al., 2021), we compared CLIP's attention alignment with human participant groups with different visual attention patterns discovered in the EMHMM analysis. We aggregated all subjects in each group instead and calculated PCC between CLIP and each group separately. To examine if CLIP's visual attention is more similar to one human group than the other, we constructed a similar linear mixed effects model with visual attention alignment as the dependent variable, participant group of comparison, condition (abstract as the baseline), and their interaction as fixed effects and image as a random effect.

Results

Concreteness Bias in CLIP and Human

The linear regression analyses revealed a robust positive relationship between word concreteness and word importance in CLIP (red line in Figure 3; $R^2 = 0.23$, $p < .001$), confirming that the concreteness bias reported in Zhao et al. (2025) persisted in the dataset sampled for the current study. In contrast, this relationship was substantially weaker in humans ($R^2 = 0.0063$, $p = .012$)

The linear mixed effect model comparing the concreteness bias effect between CLIP and humans showed a significant main effect of agent on word importance score, with CLIP assigning overall lower word importance than humans ($\beta = -1.16$, $SE = 0.10$, $p < .001$). This overall scale difference can be explained by CLIP's sparser textual attention compared with humans: while humans tended to look at most of the words in a sentence, potentially due to the sequential reading order, CLIP only attended to a few keywords (Figures 2 and 3). Importantly, we observed a significant interaction between concreteness and agent (CLIP vs. human, $\beta = 0.23$, $SE = 0.02$, $p < .001$). While the concreteness effect was significant for both agents, the bias was positive for CLIP ($\beta = 0.18$, $SE = 0.02$, $p < .001$) but negative for human ($\beta = -0.05$, $SE = 0.02$, $p = .003$), after using the mixed effects model to account for the random effect of images.

This opposite effect observed in humans may be explained by their more frequent opportunities of using and learning about abstract language during everyday interactions and communications, and by the richer, multimodal learning environment with more complex semantic information during human development, as compared with CLIP's training, which is restricted to matching static images with whole sentences through contrastive learning, where abstract concepts may be difficult to discern or infer.

Figure 3: Word importance vs. word concreteness for all words in the sampled captions (same word in different captions treated separately). Word concreteness is rated from 1 (less concrete) to 5 (more concrete). Red lines indicate linear regression results for CLIP ($R^2 = 0.23$, $p < .001$, left) and human ($R^2 = 0.0063$, $p = .012$, right).

CLIP-Human Attention Alignment

Our linear mixed effect model revealed a significant main effect of condition on CLIP-human *caption* attention alignment: the alignment was higher when they matched the images with concrete captions than with abstract captions ($\beta = 0.082$, $SE = 0.037$, $p = .028$). This finding supports our hypothesis that CLIP's information use diverges more from humans' when processing abstract language, potentially

reflecting CLIP's difficulty in learning abstract concepts purely from matching static images and whole sentences through contrastive learning, resulting in reduced attention to and weaker representations of abstract words, as compared with humans' more robust abstract language understanding.

In contrast, the linear mixed-effects model examining CLIP–human image attention alignment on *images* revealed no significant effect of condition (β = −0.003, SE = 0.022, p = .892), suggesting that CLIP-human alignment in attention to the images did not differ when matching the images with concrete or abstract captions. Thus, the difference between CLIP and humans in attention to concrete vs. abstract words during image-caption matching did not appear to be reflected in their attention patterns to the images. However, prior work has shown that humans exhibit individual differences in visual attention strategies or perceptual style, characterized as explorative versus focused, across a range of image-viewing tasks with various datasets (e.g., Hsiao et al., 2021; Qi et al., 2024). This individual difference may attenuate condition- or task-dependent effects, potentially contributing to the absence of a condition effect in the present analysis. Therefore, we proceeded to examine if such individual difference persisted in the current image-caption matching context and whether CLIP was more aligned with one group than the other.

Individual Difference in Visual Attention Pattern

Using EMHMM with co-clustering for fixations on the *image only* across all image-caption pairs, we identified two distinct eye movement pattern groups: an explorative group and a focused group (Groups E and F in Figure 4, respectively), consistent with prior findings with different image-viewing tasks. Participants in the explorative group attended to broader regions of interest centered around the described objects and the overall image, whereas participants in the focused group exhibited smaller, more spatially constrained ROIs. In addition, visual attention heatmaps of Group E have higher spatial variance, measured as the second central moment of the normalized heatmap distribution, than Group F, t(159) = 15.6, p < .001.

To examine whether participants were more likely to adopt one attention strategy (i.e., greater similarity to one pattern group) in one caption condition (concrete vs. abstract) than the other, we compared AB scale, here referred to as EF scale, across the conditions. Condition-specific EF scale values were computed by averaging across all trials within each caption condition when calculating the log-likelihoods L_E and L_F. We found that the EF scale difference between the two caption conditions was insignificant, t(59) = −1.61, p = .109, suggesting that the identified individual differences reflected stable, participant-level viewing preferences that outweigh condition-specific task demands.

To compare CLIP's alignment with different human groups, our mixed effect model revealed a significant group effect on CLIP-human *image* attention alignment: CLIP's visual attention was more similar to the Explorative group than the Focused group (β = −0.06, SE = 0.02, p = .001). In contrast, there was no significant effect of condition (β = −0.01, SE = 0.02, p = .745) nor a significant group × condition interaction (β = 0.02, SE = 0.03, p = .404). This finding was consistent with Qi et al. (2024), who showed that image classification AI (ResNet50, He et al., 2016) had higher attention alignment with human participants who adopted an explorative attention strategy than those adopting a focused strategy during image classification. Indeed, whereas human observers primarily focused on the objects mentioned in the captions, CLIP tended to attend to a broader range of image details across the entire scene (Figure 2).

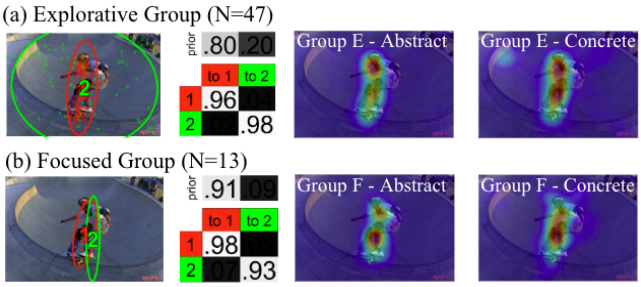

Figure 4: EMHMM with co-clustering for human's fixations on the images revealed the explorative (Group E, top) and the focused group (Group F, bottom). Ellipses show ROIs as 2-D Gaussian emissions and dots denote the raw fixations. Priors are the probabilities of fixation sequences starting from each ROI, and the matrices show the transition probabilities among ROIs. Heatmaps visualize each group's attention in the concrete and abstract conditions.

Discussion

Recent research has reported a semantic bias towards concrete words in image-text matching in the VLP model CLIP (Zhao et al., 2025). Here we examined whether a similar bias could be observed in humans and whether CLIP and humans differed in this semantic bias. Consistent with Zhao et al. (2025), we found a bias toward concrete words when matching captions with images in CLIP, as demonstrated by the positive correlation between word concreteness and word importance score (Figure 3). This bias may arise from the contrastive learning adopted by CLIP, which aligns whole images with complete captions, rather than explicitly grounding individual phrases in corresponding image regions. Indeed, when comparing the image attention of CLIP and humans (Figure 2), while both CLIP and humans showed high attention on the word "skateboarder", CLIP only attended to the skateboard on the image, whereas humans attended to the person on top of the skateboard. One possible explanation is that most images paired with the captions that include the word "skateboarder" in the dataset consistently have the skateboard object. Therefore, when CLIP learned to match those images with the whole captions, it was encouraged to rely on the skateboard visual cue to make the association, unlike humans who understand the reference of "skateboarder" to the corresponding person.

Besides the term-level concept grounding, sentence-level compositional and contextual dependencies are critical when

humans interpret and understand abstract concepts (Morid et al., 2020). To probe CLIP's understanding of compositional languages, Zhao et al. (2025) evaluated the model on visual reasoning captions from CLEVR (Johnson et al., 2017) and found that CLIP performed particularly poorly on understanding relative attributes such as size or spatial relations (e.g., "small object", "left object"), compared with attributes defined by absolute labels such as color and shape (e.g., "red object"). Lewis et al. (2024) also found that CLIP performed well on understanding single adjective–noun compositions but showed limitations on compositional tasks that require binding variables such as shape-color composition and relative phrase composition. How to improve CLIP's performance on the understanding of compositionality is currently an active research area (e.g., Yuksekgonul et al., 2022; Kang et al., 2025).

When we compared CLIP and humans on the association between word concreteness and word importance score (Figure 3), our mixed effects modeling results showed significantly different associations in CLIP and in humans: after accounting for the random effect of images, humans demonstrated a weak negative association, suggesting a bias of attending to abstract words more than concrete words, in contrast to CLIP's bias towards concrete words. Human's stronger attention on abstract words than CLIP may arise from the rich opportunities to develop solid understanding of abstract concepts during development. Unlike CLIP, which was trained solely on image-text matching and showed limited emergent ability to understand abstract language, humans' acquisition of abstract concepts is not confined to a single task or to the visual and language modalities alone. Instead, abstract concept learning in humans is supported by multiple cognitive systems, such as metacognition, interoception, and social cognition, as well as their interactions with the language system, which allow us to ground abstract concepts such as mental states, emotions, and social concepts (Borghi et al., 2020).

One factor that could potentially bias attention in both CLIP and humans is word frequency. For stochastic language models such as CLIP, which learn distributed representations of word meaning from statistical regularities in large training corpora, frequent words tend to acquire more robust and stable representations and be overrepresented than rare words (Algayres et al., 2025). In humans, word frequency is also known to modulate lexical processing, with more frequent words processed faster and with fewer fixations than less frequent words (Brysbaert et al., 2018). However, Zhao et al. (2025) showed that although CLIP assigns greater attention to concrete words, these words are in fact less frequent than abstract words in the training corpus, ruling out word frequency as a confounding factor underlying CLIP's concreteness bias. In the current experiment, we further controlled word frequency to minimize its influence on human attention on abstract vs. concrete words. Another factor that may influence human attention is word processing difficulty. Because abstract words are harder to interpret, humans may allocate more fixations to them during reading. In any case, increased attention to abstract words as compared with CLIP reflects humans' deeper semantic processing on those words, in contrast to a tendency to skip abstract information observed in CLIP.

Our analysis of CLIP-human attention alignment revealed that CLIP's attention to *captions* was more closely aligned with humans when processing concrete captions than the abstract captions. This pattern suggests that CLIP's information use diverges more from humans when processing abstract language, likely reflecting limitations of contrastive learning in forming robust representations of abstract concepts. In contrast, for *image* attention, their alignment did not differ between caption conditions, since human participants' attention to the images was primarily governed by stable individual perceptual styles rather than by task demands associated with matching concrete versus abstract captions. Using EMHMM with co-clustering, we discovered the explorative and focused attention strategies when matching images with captions. Such perceptual style difference is consistent with prior findings from other visual tasks, including passive viewing and image classification. Interestingly, CLIP's image-level attention aligned more closely with the explorative than the focused group, a pattern also observed in image classification tasks comparing humans with the convolutional neural network ResNet50 (Qi et al., 2024; He et al., 2016). This convergence suggests a characteristic information-use tendency in deep neural networks toward more distributed visual attention.

While the present study focused on differences in *attention patterns* between CLIP and humans while controlled for image–text matching performance, future research could directly examine whether these two agents *rate* concrete and abstract image–caption pairs differently. Such evaluations may provide further insight into whether CLIP's output is consistent with human expectations, which may impact the performance of downstream applications.

In conclusion, we found that CLIP exhibited a semantic bias toward concrete words during image–text matching, whereas humans attended to abstract words more instead. CLIP's concreteness bias likely results from the contrastive training, which potentially limited its ability to develop a grounded understanding of concepts and capture contextual dependencies between images and captions. While CLIP's caption attention aligned more closely with humans for concrete than for abstract captions, alignment in image attention did not differ between caption conditions, as humans' image attention was primarily driven by stable individual perceptual styles, favoring either a focused or explorative viewing strategy, rather than by task demands associated with different types of captions to match. Our findings thus revealed important differences in information processing mechanisms for image-text matching between humans and CLIP, with important implications not only for our understanding of human cognition and AI, but also for potential ethical issues resulting from such misalignment.

Acknowledgments

This study was supported by Research Grant Council of Hong Kong (AoE/E-601/24-N) and Research Grant from City University of Hong Kong (Project No. 7030010). We thank Yuhang She for her help in data collection.

References

Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197.

Algayres, R., Saint-James, C.-É., Luthra, M., Shen, J., Lin, D., Benchekroun, Y., Moritz, R., Pino, J., & Dupoux, E. (2025). LongTail-Swap: benchmarking language models' abilities on rare words. *arXiv preprint arXiv:2510.04268.*

Bleasdale, F. A. (1987). Concreteness-dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(4)*, 582-594.

Borghi, A. M., Flumini, A., Cimatti, F., Marocco, D., & Scorolli, C. (2011). Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Frontiers in psychology, 2*, 15.

Borghi, A. M. (2020). A future of words: Language and the challenge of abstract concepts. *Journal of Cognition, 3(1)*, 42.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current directions in psychological science, 27*(1), 45-50.

Changpinyo, S., Sharma, P., Ding, N., & Soricut, R. (2021). Conceptual 12m: Pushing web scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3558-3568).

Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 397-406).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hsiao, J. H., Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye movement analysis with hidden Markov models (EMHMM) with co-clustering. *Behavior Research Methods, 53*(6), 2473–2486.

Hsiao, J. H., & Chan, A. B. (2023). Towards the next generation explainable AI that promotes AI-human mutual understanding. *NeurIPS XAIA.*

Hsiao, J. H. (2026). Comparability between AI and human cognition and its role in psychological research and ethical AI. *British Journal of Psychology*.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2901-2910).

Kang, R., Song, Y., Gkioxari, G., & Perona, P. (2025). Is CLIP ideal? No. Can we fix it? Yes!. *arXiv preprint arXiv:2503.08723.*

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*(2), 325-343.

Lewis, M., Nayak, N., Yu, P., Merullo, J., Yu, Q., Bach, S., & Pavlick, E. (2024). Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1487-1500).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014, PT V* (Vol. 8693, Issue 13th European Conference (ECCV 2014), pp. 740-755). Springer International Publishing.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2022). Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, *508*, 293-304.

Morid, M., Griffiths, S. W., & Zamuner, T. S. (2020). The Representation of Concrete Versus Abstract Words: An Eye-tracking Study. *Journal of the Canadian Linguistic Association*

Paivio, A., Yuille, J. C., & Madigan, S. A. (1986). Concreteness, imagery and meaningfulness values for 925 words. *Journal of experimental psychology, 76*(1, Pt.2), 1–25.

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2024). Explanation Strategies in Humans vs. Current Explainable AI: Insights from Image Classification. *British Journal of Psychology, 00,* 1-24.

Schwanenflugel, P. J., & Stowe, R. W. (1989). Context Availability and the Processing of Abstract and Concrete Words in Sentences. *Reading Research Quarterly, 24(1)*, 114–126.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology Section A, 57*(1), 165-182.

Yang, S. C. H., Vong, W. K., Sojitra, R. B., Folke, T., & Shafto, P. (2021). Mitigating belief projection in explainable artificial intelligence via Bayesian teaching. *Scientific reports, 11*(1), 9863.

Yang, S. C. H., Folke, N. E. T., & Shafto, P. (2022). A psychological theory of explainability. In *International Conference on Machine Learning* (pp. 25007-25021). PMLR.

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. H. (2023). Humans vs. AI in detecting vehicles and humans in driving scenarios. *Proceedings of the Annual Conference of the Cognitive Science Society, 45,* 1832-1839.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it?. *arXiv preprint arXiv:2210.01936.*

Zhao, C., Wang, K., Zeng, X., Zhao, R., & Chan, A. B. (2024). Gradient-based visual explanation for transformer-based clip. In: *International Conference on Machine Learning* (pp. 61072-61091).

Zhao, C., Wang, K., Hsiao, J. H., & Chan, A. B. (2025). Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP. *arXiv preprint arXiv:2502.18816.*

Zhou, C., Loy, C. C., & Dai, B. (2022). Extract free dense labels from clip. In *European conference on computer vision* (pp. 696-712).