OPEN MIND
Discoveries in Cognitive Science

an open access journal

REPORT

DeepMeaning: Estimating and Interpreting Scene Meaning for Attention Using a Vision-Language Transformer

Taylor R. Hayes[1] and John M. Henderson[1,2]

[1]Center for Mind and Brain, University of California, Davis
[2]Department of Psychology, University of California, Davis

Keywords: scene perception, attention, semantics, meaning maps, vision-language transformer

ABSTRACT

Humans rapidly process and understand real-world scenes with ease. Our stored semantic knowledge gained from experience is thought to be central to this ability by organizing perceptual information into meaningful units to efficiently guide our attention. However, the role stored semantic representations play in attentional guidance remains difficult to study and poorly understood. Here, we apply a state-of-the-art vision-language transformer trained on billions of image-text pairs to help advance our understanding of the role local meaning plays in scene guidance. Specifically, we demonstrate that this transformer-based approach can be used to automatically estimate local scene meaning in indoor and outdoor scenes, predict where people look in these scenes, detect changes in local semantic content, and provide multiple avenues to model interpretation through its language capabilities. Taken together, these findings highlight how multimodal transformers can advance our understanding of the role scene semantics play in scene attention by serving as a representational framework that bridges vision and language.

Citation: Hayes, T. R., & Henderson, J. M. (2025). DeepMeaning: Estimating and Interpreting Scene Meaning for Attention Using a Vision-Language Transformer. *Open Mind: Discoveries in Cognitive Science*, *9*, 1020–1036. https://doi.org/10.1162/opmi.a.6

DOI:
https://doi.org/10.1162/opmi.a.6

Received: 17 June 2024
Accepted: 21 May 2025

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Taylor R. Hayes
trhayes@ucdavis.edu

The MIT Press

INTRODUCTION

Semantic knowledge is central to how we perceive and make sense of the complex visual world around us (Murphy, 2004; Reilly et al., 2025). While semantic representations are commonly thought of in linguistic terms as the mapping of a word or phrase to a specific object or concept, semantic representations also organize perceptual information into meaningful units that help to efficiently guide our attention in scenes (Henderson, 2007, 2011). Therefore, improving our understanding of the interplay between semantic representations and attention in scenes has the potential to have both broad theoretical impact and to advance a variety of nascent technologies which require rapid scene understanding (e.g., autonomous cars and other agents). While scene semantics are difficult to study and remain poorly understood, recent advancements have made their study more tractable (Hayes & Henderson, 2021b; Henderson & Hayes, 2017). Here we take another step toward understanding semantics in scenes by applying a state-of-the-art vision-language transformer (Yu et al., 2022) to accurately estimate local scene meaning, predict scene attention, and provide a direct route to interpreting estimates of local scene meaning.

Cognitive guidance theory is the theoretical framework anchoring our work (Henderson, 2003, 2011). Under this view, semantic knowledge stored in memory 'pushes' our attention toward scene regions that are recognizable, informative, and relevant to our current goals (Biederman, 1972; Henderson & Hollingworth, 1999; Land & Hayhoe, 2001; Potter, 1975; Wolfe & Horowitz, 2017). That is, where we look in scenes is primarily driven by semantic representations that guide our attention toward meaningful scene regions. There is a long history of evidence supporting the relationship between semantic properties and attention in scenes (Antes, 1974; Buswell, 1935; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Torralba et al., 2006; Williams & Castelhano, 2019; Yarbus, 1967), including demonstrations that scene semantics often supplant nonsemantic visually salient scene regions (Võ et al., 2019; Williams & Castelhano, 2019; Wu et al., 2014). However, one major limitation of much of this earlier work is that it often focused on isolated object-scene semantic relationships (e.g., swapping an octopus and a tractor in an underwater and farm scene respectively; Loftus & Mackworth, 1978). While these discrete semantic manipulations were important in establishing a causal relationship between scene semantics and attention, they do not tell us much about the overall role of semantic guidance of attention in scene understanding (Henderson & Hayes, 2017).

Two recent studies introduced different approaches to studying the effects of scene semantics globally across entire scenes: meaning maps (Henderson & Hayes, 2017) and concept maps (Hayes & Henderson, 2021b). Meaning maps use human raters to estimate a given semantic feature at each location in the scene. Specifically, each scene (Figure 1a) is broken into small circular image patches at two spatial scales (Figure 1b), and then participants rate a random subset of these image patches based on a given semantic instruction (e.g., meaningful, informative and recognizable; Henderson & Hayes, 2017). These ratings are then combined back into their respective position to form a map of local scene meaning. Local scene meaning has repeatedly been shown to be one of the strongest predictors of where people look in scenes regardless of the viewing task (See Henderson et al., 2019, for review). In addition to local meaning maps, we also developed a separate language-based approach using a vector space semantic model called ConceptNet Numberbatch (Hayes & Henderson, 2021b). ConceptNet Numberbatch derives the semantic relationships between words based on regularities

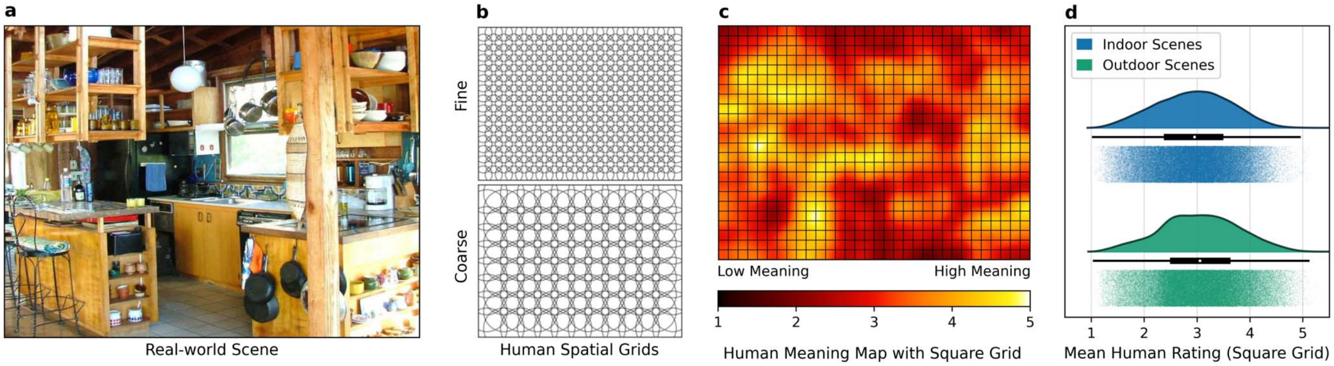

Figure 1. Meaning mapping and input/target preprocessing. A meaning map for each scene (a) is built by breaking each scene into circular patches at two spatial scales (b) and then having humans rate the patches. The human patch ratings are then recombined to generate a scene meaning map (c). To train DeepMeaning, each scene image (a) and meaning map (c) were broken into patches using a square grid (c). The square scene image patches served as the input to the pretrained Vision transformer (ViT) of the Contrastive Captioner (CoCa) while the average meaning map value of each square region served as the target value to be predicted. Raincloud plots of the distribution of the meaning target values for indoor and outdoor scenes were normally distributed (d).

in almost a trillion words of written text and crowd-sourced basic knowledge about the world (Günther et al., 2019). The semantic representations from ConceptNet can then be mapped back onto the objects in a scene to form a 'concept map' that reflects how semantically related each object is to the rest of the scene, which is also strongly associated with scene attention (Hayes & Henderson, 2021b).

Therefore, meaning maps and concept maps each approach scene semantics from a different angle. Meaning maps are constructed by filtering a visual stimulus through the cognitive system of human raters to estimate semantic properties in scenes (e.g., local meaning, Henderson & Hayes, 2017; graspability, Rehrig et al., 2020; interaction, Rehrig et al., 2022), while concept maps (Hayes & Henderson, 2021b) are non-visual, building semantic representations based entirely on regularities in human-generated language. However, humans often acquire semantic knowledge through an interplay of visual and language experience (Clarke & Tyler, 2015; Ralph et al., 2017), so scene semantics may best be understood within a computational framework that forms a multimodal mapping between vision and language.

Here we apply just such a framework, a state-of-the-art Contrastive Captioner (CoCa) which serves as a foundational vision-language representational model (Yu et al., 2022). While transformers have played a large role in natural language processing, it is only recently that transformers have been generalized to also include visual and multimodal vision-language domains (Dosovitskiy et al., 2021; Vaswani et al., 2017; Yu et al., 2022). CoCa in particular recently introduced a unique architecture that unifies many of the strengths of previous transformer architectures (i.e., single encoder, dual encoder, and encoder-decoder), allowing CoCa to learn aligned unimodal text and image embeddings as well as a fused multimodal image-text representational space (Yu et al., 2022). It is this unique ability that allows CoCa to learn very general representations that achieve state-of-the-art performance across virtually every major vision, language, and multimodal benchmark (Yu et al., 2022). CoCa has proven successful in the large and growing transformer benchmarking literature. Importantly, our primary goal here is not to further benchmark CoCa against other models, but rather to test whether the general CoCa feature space can be used to both estimate and interpret local scene meaning. That is, we apply CoCa as a tool to investigate the role of human conceptual knowledge in the interplay between attention and scene understanding.

In the present study, we used the feature space of CoCa to train a linear model to estimate local meaning (Figure 2), predict attention (Figure 3), and interpret local scene meaning (Figures 4 and 5) in an application we call 'DeepMeaning'. The overview of how DeepMeaning estimates local scene meaning is shown in Figure 2a, and can be broadly split into a *feature extraction stage* and a *leave-one-scene-out cross-validation stage*. In the feature extraction stage, we take the CoCa model pretrained on more than 2 billion unique image-text pairs (Figure 2a, purple) and use it to generate CoCa features for each local scene region by breaking each scene into smaller patches using a square grid (Figure 2a, white). Then, we train a linear model (Figure 2a, red) for indoor scenes and a linear model for outdoor scenes where we use these general CoCa features for the scene patches as predictors to estimate local meaning using a leave-one-scene-out training and testing procedure (Figure 2a, grey). Indoor and outdoor scenes were modeled separately because there is evidence that they are behaviorally (Torralba et al., 2006) and neurally distinct (Henderson et al., 2007). A direct comparison of the linear indoor and outdoor weights showed they only shared about 11% of their variance, providing additional justification for separate indoor and outdoor models. Using this general procedure, we evaluated DeepMeaning based on meaning recovery (local patches and full scenes), attention prediction,

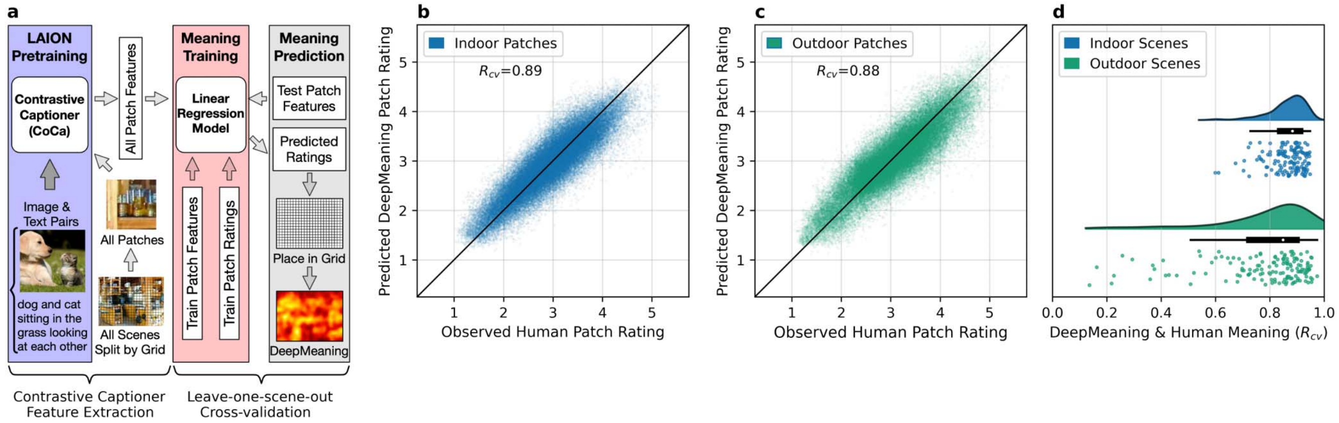

Figure 2. DeepMeaning overview and meaning recovery results. DeepMeaning combines the features from a Contrastive Captioner (CoCa) transformer pretrained on billions of image-text pairs with a linear model to predict patch meaning ratings (a). The scatterplots (b-indoor, c-outdoor) show DeepMeaning's patch-level meaning prediction relative to human meaning ratings where each dot in the plot represents an individual scene patch. The raincloud plots (d) show the distribution of the correlations between the DeepMeaning predicted meaning map and the ground truth human meaning map, where each dot represents a left-out indoor or outdoor scene.

ability to detect changes in semantic content, and model interpretability. These four criteria can broadly be seen as evaluating the two major goals of the present manuscript: showing that a multimodal transformer can serve as a tool that provides image-computable local scene meaning ratings with human-like characteristics, and demonstrating that the aligned vision-language feature space can be used to directly interpret estimates of local scene meaning.

While the practical benefit of an image-computable method to estimate local scene meaning without collecting thousands of human patch ratings is clear, the interpretive ability offered by a vision-language model is equally important. In traditional vision-only approaches, a deep neural network (convolutional neural network or vision transformer) is trained to map a visual input to an output (e.g., object categories). However, between the visual input and the model output are dozens of hidden layers composed of thousands of units that make it notoriously hard to understand exactly what features these models rely on to make their estimates (Bowers et al., 2022). And while in some domains all that matters is how well a model performs, in

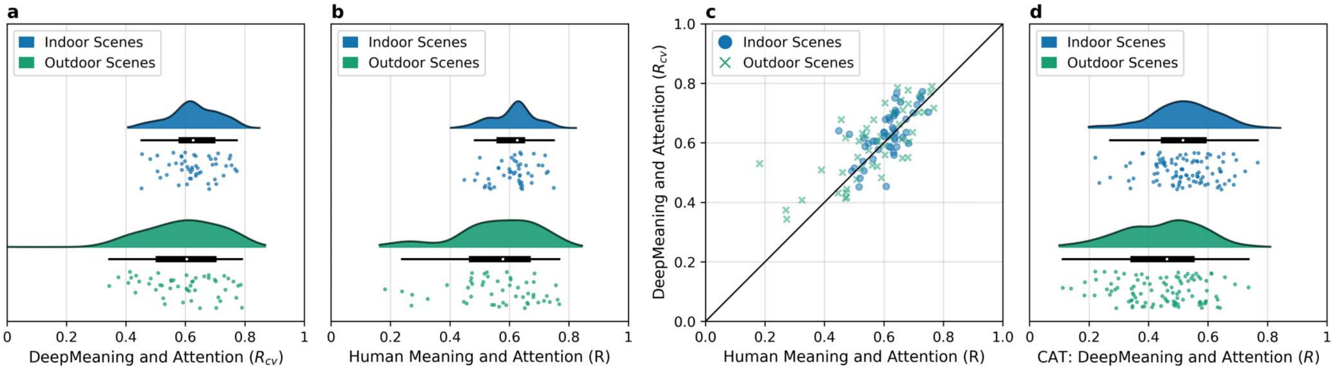

Figure 3. DeepMeaning maps transfer to predict scene attention just like human meaning maps. Raincloud plots show that DeepMeaning maps (a) and human meaning maps (b) both correlate strongly with scene attention. Moreover, the correlations between human meaning maps and attention and DeepMeaning maps and attention were very similar (R_{cv} = 0.90) scene to scene (c). Finally, we applied DeepMeaning to an additional scene dataset with fewer observers (CAT2000) and replicated a strong correlation between DeepMeaning maps and scene attention (d) for both indoor and outdoor scenes.

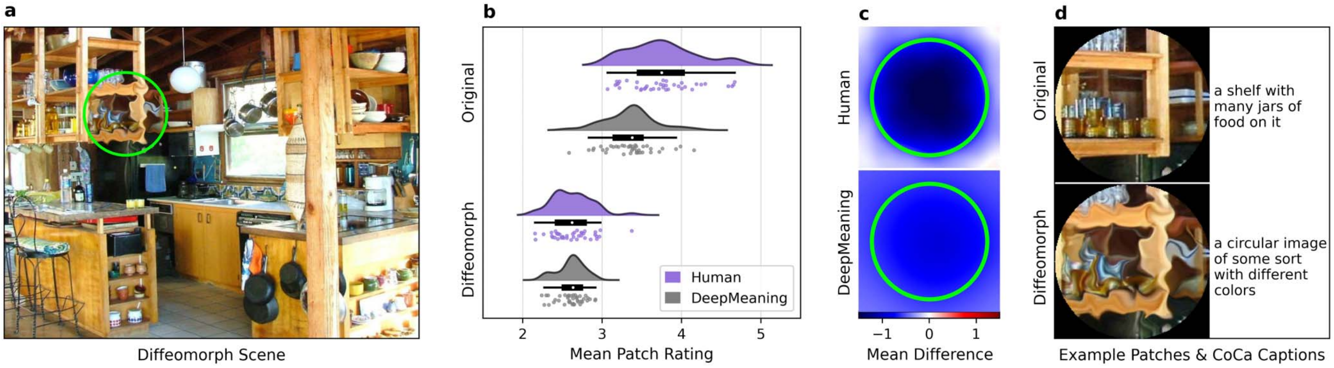

Figure 4. DeepMeaning detects the removal of semantic content. We applied DeepMeaning to 40 scenes where semantic content was removed while preserving image features via a diffeomorphic transform (a, d). DeepMeaning passed this semantic content test just like human meaning maps, showing a large decrease in meaning value for the diffeomorphed scene region relative to the original non-diffeomorphed scene region (b). The mean difference in rating is shown spatially by averaging across the critical region (c). Additionally, we used CoCa to decode captions for the original and diffeomorphed image patch (d), revealing that CoCa could no longer identify a mapping between semantically meaningful objects, offering a human-interpretable explanation for the drop in meaning values.

cognitive science, being able to interpret how a model makes its predictions is in many ways just as important as the estimates themselves. Simply put, black boxes are not conducive to guiding cognitive theory no matter how well they can map inputs to outputs. Multimodal transformers offer a potential solution by aligning visual and linguistic features, enabling direct interpretation of model behavior through language. Specifically, this vision-language alignment enables two direct paths to interpretation: generating natural language descriptions of local scene regions, and using contrastive prompts to test how model estimates align with specific, user-defined semantic features.

METHODS

Contrastive Captioner (CoCa)

We selected the Contrastive Captioner (CoCa) model (Yu et al., 2022) as our embedding space because its vision-language architecture generates general features well-suited for downstream tasks, while also providing direct avenues for interpretation. In our investigation, CoCa embeddings also outperformed two vision-only transformer baselines: the Self-distillation with No Labels (DINO, Caron et al., 2021) model and the Masked Autoencoder (MAE, He et al., 2021) model–in both patch-level and scene-level meaning prediction (see supplementary materials). These results suggest that language-aware embeddings enhance local meaning prediction, further validating our choice of CoCa for our embedding space.

We used the OpenClip Contrastive Captioner (CoCa) implementation (coca_ViT-L-14 with the *mscoco_finetuned_laion2b_s13b_b90k* pretrained weights, Ilharco et al., 2021) based on the original CoCa model by Yu et al. (2022). The OpenClip CoCa model was pretrained on 13 billion samples from the LAION-2B dataset using a batch size of 90,000, a learning rate of 1e-3, and a cosine decay learning rate schedule (Schuhmann et al., 2022). These weights (728 dimensional feature space) were then finetuned using the Microsoft COCO dataset (Lin et al., 2014) using a batch size of 128, a learning rate of 1e-5, and a cosine learning rate schedule (Schuhmann et al., 2022). The LAION-2B dataset is the English subset of the larger multilingual LAION-5B dataset. The LAION-2B dataset is an open dataset for model training that contains 2.32 billion image-text pairs (Schuhmann et al., 2022).

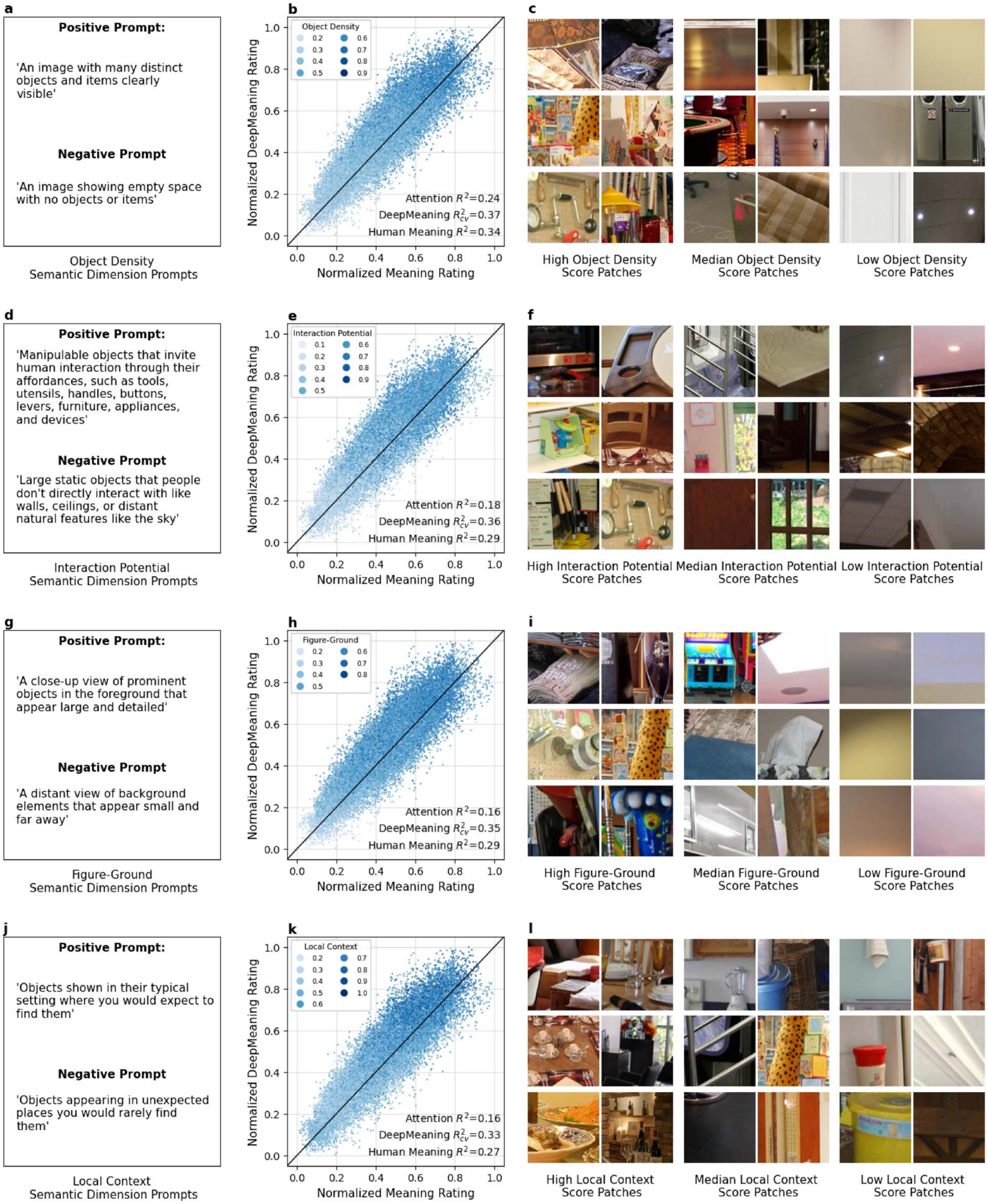

Figure 5. Semantic projection analysis using contrastive prompts. (a, d, g, j) Contrastive prompts were used to define concrete semantic directions (e.g., object density). (b, e, h, k) Scatter plots show the relationship between normalized semantic scores and meaning ratings, with points colored by their semantic score (darker blue indicates higher scores). R^2 values quantify variance explained in attention, DeepMeaning ratings, and human meaning ratings. (c, f, i, l) Representative patches with high, median, and low semantic scores, demonstrating how the semantic scoring aligns with visual content.

DeepMeaning

Scene Preprocessing. Each scene and its corresponding meaning map were split into 128 × 128 pixel square patches with 73% overlap (Figure 1c). Patch overlap was used based on previous meaning map work showing a benefit to overlap for recovering known visual features (Henderson & Hayes, 2018). Each square scene image served as an input to the vision transformer (ViT) component of CoCa for feature extraction. The meaning value for each square scene region was computed as the average across its location in the corresponding human meaning map and served as the target value to be predicted by DeepMeaning (Figure 1d).

Architecture. DeepMeaning is composed of two components: a pretrained Contrastive Captioner (CoCa) transformer that is used as a feature extractor and a linear model that is trained to use these extracted features to predict human scene meaning ratings. Specifically, the pretrained weights (728 dimensional feature space) learned by the Contrastive Captioner by training on the LAION-2B dataset were frozen, and then used to extract general features from each square scene image patch (Figure 2a). The extracted image patch features and their corresponding meaning ratings were then used to train a linear model to predict meaning ratings for indoor and outdoor scene patches separately using a leave-one-scene-out cross-validation procedure.

Leave-one-scene-out cross-validation procedure. The leave-one-scene-out train/test cross validation procedure was used to estimate the generalization performance of DeepMeaning on our training scene set ($N = 282$). In this procedure, the linear model component of DeepMeaning was trained on all scenes but one, and then the trained linear model weights were frozen and used to predict the meaning values for the left-out-scene image patches. This procedure was done separately for indoor ($N = 139$) and outdoor scenes ($N = 143$) producing a separate set of linear weights for indoor and outdoor scenes.

DeepMeaning Ensemble Indoor and Outdoor Models. To mitigate overfitting, we employed model averaging across all cross-validation folds to generate the final DeepMeaning linear models for indoor and outdoor scenes. For each leave-one-scene-out fold, we saved the model weights and intercept. The final ensemble models were then created by averaging these parameters across all folds—139 folds for indoor scenes and 143 folds for outdoor scenes. This ensemble approach produced two robust models capable of predicting meaning in novel scenes: one optimized for indoor environments and another for outdoor environments.

The ensemble indoor and outdoor DeepMeaning linear models were validated on a dataset not used to train the models (CAT2000 dataset, Borji & Itti, 2015). The ensemble models ('DeepMeaning_indoor_ensemble.pkl' and 'DeepMeaning_outdoor_ensemble.pkl') are the model weights used in our provided Python code (see batch_deep_meaning.py) to allow users to generate DeepMeaning maps for novel indoor and outdoor scenes.

The decision to fit separate linear models for indoor and outdoor scenes was made a priori based on previous evidence that indoor and outdoor scenes are distinct (Henderson et al., 2007; Torralba et al., 2006). In addition, a post hoc comparison of the DeepMeaning ensemble indoor and outdoor weights provided converging evidence for fitting separate indoor and outdoor models. The squared correlation showed the indoor and outdoor model ensemble weights only shared 11% of their variance ($R^2 = 0.11$) and were significantly different (paired samples t-test of the absolute difference between indoor and outdoor DeepMeaning weights; $t(767) = 35.95$, $p < 0.001$, 95% CI [0.63, 0.71]).

Meaning Map Data

Participants. University of California, Davis undergraduate students (*N* = 1149) with normal or corrected-to-normal vision participated in the meaning rating study in exchange for course credit. All participants were naïve concerning the purposes of the experiment and provided verbal or written informed consent as approved by a University Institutional Review Board. All experiments were performed in accordance with relevant guidelines and regulations.

Stimuli. 282 real-world scene images were meaning mapped. The 282 scenes consisted of a mix of indoor (139) and outdoor (143) scenes and included scenes from 100 unique scene categories (e.g., kitchen, office, park, street, etc.).

Meaning Mapping Procedure. Meaning maps were generated for each scene using the same meaning mapping procedure and rating instructions (see https://osf.io/654uh/ for the code and complete rating instructions) as Henderson and Hayes (2017). Specifically, a meaning map was created for each scene by dividing the entire scene into a dense array of overlapping circular patches at a fine and coarse spatial scale (see Figure 1b). Human raters then provided ratings of 300 random fine or coarse scene patches based on how informative or recognizable they thought they were on a 6-point Likert scale (Henderson & Hayes, 2017; Mackworth & Morandi, 1967). Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each unique patch was rated by three unique raters.

A meaning map (Figure 1c) was generated for each scene by averaging the patch rating data at each spatial scale separately and then averaging the spatial scale maps together.

Estimate of Human Rater Noise Ceiling. An estimate of the noise ceiling was computed to estimate how well DeepMeaning could potentially perform given the noise in human ratings of meaning. To perform this estimate we compared meaning maps from 40 scenes (34 indoor, 6 outdoor) from two different groups of raters in two previous studies (Hayes & Henderson, 2022; Henderson & Hayes, 2017). These correlations were performed by excluding the diffeomorphed region (about 4 percent of scene) in each scene and then computing the correlation for the remaining 96% of the scene. Since the correlations were not normally distributed, we used bootstrapping to estimate the 95% confidence intervals around the mean correlation coefficient (10000 bootstrap samples). This provided a more robust estimate of the noise ceiling that does not assume normality in the sampling distribution.

Eyetracking Datasets

Dataset 1: Internal. Eye tracking data from a large dataset were used to verify DeepMeaning's ability to transfer to predict scene attention (Cronin et al., 2020; Hayes & Henderson, 2021b). This global dataset contained 49 indoor scenes and 51 outdoor scenes spanning 100 unique scene categories. Each scene was viewed for 12 seconds by 100 observers. Observers performed a scene memorization task for half the scenes and an aesthetic judgment task for the other half of the scenes. Task instruction order was counterbalanced across observers and scenes such that all observers viewed all 100 images and each of the 100 images appeared equally under the two viewing task conditions. Since viewing task is not relevant for the question at hand, we performed our analysis below pooled across all the data. Observer eye movements were recorded using an EyeLink 1000+ tower-mount eye tracker (spatial resolution 0.01°) sampling at 1000 Hz SR Research (2010). Participants sat 85 cm away from a 21" monitor and viewed scenes that subtended approximately 27° × 20° of visual angle.

Dataset 2: External. Additionally, one hundred indoor and outdoor scenes from the CAT2000 benchmark eye tracking dataset served as an additional replication of DeepMeaning's ability to estimate local meaning that transfers to predict scene attention in datasets with a smaller number of observers and shorter viewing duration (Borji & Itti, 2015). Each scene in the CAT2000 dataset was freely-viewed by 24 observers for 5 seconds while their eye movements were recorded using an EyeLink 1000 eye tracker (SR Research, 2010).

Diffeomorph Data

The diffeomorph scene set from Hayes and Henderson (2022) was used to assess whether DeepMeaning could successfully detect the local removal of semantic content from a scene. The diffeomorph dataset contained 40 scenes in two conditions: diffeomorphed (Hayes & Henderson, 2022) and original (data from Henderson & Hayes, 2017). In the diffeomorph condition, a diffeomorphic transformation (max distortion = 15, step number = 10, Stojanoski & Cusack, 2014) was applied to one local circular region (205 pixel diameter) and then blended into each scene using a radially symmetric linear alpha ramp (15 pixel diameter) to remove the semantic content from the target region while preserving its image features (Hayes & Henderson, 2022). In the original condition, the scenes were presented unaltered. Human meaning ratings were collected for both the original scenes (N = 164) and the diffeomorphed scenes (N = 164) using the same *Meaning Mapping Procedure* described above.

Interpretation Analyses

Caption analysis. To interpret DeepMeaning's outputs, we analyzed CoCa-generated captions for scene patches before and after diffeomorphic transformation. This approach provided natural language descriptions of how semantic content changed following spatial transformation. We input both original and diffeomorphed circular images as square patches with black boundaries to accommodate CoCa's square input requirement (Figure 4d). For each scene (N = 40) and condition (original and diffeomorph), we generated captions for the target scene region using the same CoCa caption generation parameters (5% quantile token generation, temperature = 1.5, and repetition penalty = 22, Ilharco et al., 2021).

Our caption analysis examined two aspects: (1) caption accuracy – whether captions correctly described the original patch content (yes or no), and (2) object count changes – the percentage decrease in identified objects between original and diffeomorph captions. The authors performed the first analysis through direct visual comparison of patches and captions, and the second analysis by counting the number of objects in each caption. These metrics quantified the general loss of semantic content and specific changes in object recognition following diffeomorphic transformation, demonstrating how visual-language models' caption generation ability can provide interpretable insights into local semantic changes.

Semantic Projection Analysis. We developed a contrastive prompting method using CoCa's vision-language embeddings to analyze how specific semantic dimensions contribute to local meaning and attentional guidance. For four dimensions (object density, interaction potential, figure-ground organization, and local context consistency), we created paired prompts describing the presence versus absence of each semantic attribute. Converting these prompt pairs to CoCa text embeddings and calculating their difference vectors defined semantic directions in the embedding space. Projecting each patch's image embedding onto these directions yielded scalar scores quantifying each individual patch's alignment with each semantic dimension.

We analyzed 49 indoor scenes from eye tracking dataset 1 to allow a comparison to DeepMeaning estimates, human meaning ratings, and attention. Indoor scenes were chosen for this demonstration since they are less noisy than outdoor scenes and contain more semantic content. For each semantic dimension of interest, we compared the semantic patch projection scores against three patch measures: DeepMeaning estimate, mean human meaning rating, and mean fixation density. We calculated both total variance explained and unique contributions of each predictor, revealing how concrete semantic attributes influence local meaning estimates and attention allocation.

Dominance analysis and its interpretation. We performed a dominance analysis (Azen & Budescu, 1993, 2003) to quantify each semantic direction's unique contribution to the total variance explained in DeepMeaning estimates, human meaning ratings, and attention (since the semantic directions were not orthogonal, see supplementary materials). Dominance analysis estimates the relative importance of individual predictors in multiple regression by calculating the average increase in R-squared when a predictor is added to all possible subset regression models. In our analysis, for each contrastive semantic direction, we computed its average contribution to R^2 across all possible subsets of predictors. The contribution for each subset size was weighted by the number of combinations containing that predictor. These contributions were then normalized and multiplied by 100 to yield percentages of the total explained variance for each semantic direction to DeepMeaning estimates, human meaning ratings, and attention (e.g., in Table 1, object density accounted for 25.3% of the total variance in DeepMeaning ratings).

The dominance analysis results are most interpretable for DeepMeaning because its maximum explainable variance is $R^2 = 1.0$ because DeepMeaning's estimates and the semantic projections are both derived directly from the same CoCa embedding space. In contrast, human meaning ratings have a noise ceiling of approximately $R = 0.87$ due to inter-rater variability, while attention data is similarly limited by individual viewing patterns, task dependencies, and eye tracking measurement noise (a leave-one-subject out cross validation of fixation density provided an estimated maximum of $R^2 = 0.56$). For this reason, DeepMeaning R^2_{cv} provides the clearest measure of each semantic direction's contribution to local meaning, while the human meaning and attention data are slightly less interpretable (given their maximum R^2 are not 1.0) but highlight that the semantic projection scores do indeed transfer to the human behavioral rating and attention data as we would expect given their strong correlations with DeepMeaning estimates.

Table 1. Semantic Projection Analysis and Dominance Analysis Results

		Dimension R^2				Dominance Analysis			
	Total R^2	OD	Int	FG	Ctx	OD	Int	FG	Ctx
DeepMeaning	0.555	0.372	0.360	0.349	0.332	25.3	25.1	24.7	24.9
Human Meaning	0.465	0.336	0.293	0.289	0.274	26.2	24.6	24.5	24.7
Attention	0.294	0.236	0.178	0.157	0.165	28.3	24.4	23.1	24.2

Note: OD = Object Density; Int = Interaction; FG = Figure-ground, Ctx = Context.

DeepMeaning theoretical max $R^2_{cv} = 1.0$; Human Meaning ceiling $R^2 = 0.76$; Attention ceiling $R^2 = 0.56$.

RESULTS

We first tested how well DeepMeaning could recover local scene meaning compared to human raters (Figure 2). Using the leave-one-scene-out cross-validation procedure (Figure 2a), DeepMeaning showed excellent recovery at both the individual patch-level (indoor R_{cv} = 0.89, outdoor R_{cv} = 0.88; Figures 2b and 2c) and for scene-level maps (indoor R_{cv} = 0.87, bootstrap 95%CI [0.85, 0.88]; outdoor R_{cv} = 0.78, bootstrap 95%CI [0.75, 0.80], Figure 2d). To place DeepMeaning's scene-level performance in context relative to human raters, when two different groups of human raters (Hayes & Henderson, 2022; Henderson & Hayes, 2017) rated 40 scenes (34 indoor, 6 outdoor) the scene-level correlation observed between the two rater groups was R = 0.87 (bootstrap 95%CI [0.85, 0.89]), which suggests DeepMeaning is performing within or very close to the noise ceiling of human raters. Similar to human raters, indoor scenes were more consistently rated by DeepMeaning than outdoor scenes (t_{Welch}(180.03) = 5.46, p < 0.001, t-distribution 95%CI [0.06, 0.12]), which is reflective of noisier human meaning ratings in outdoor scenes compared to indoor scenes (Henderson & Hayes, 2017, 2018).

Next we evaluated whether DeepMeaning maps were strongly associated with where people looked in each scene (Figure 3). Specifically, we correlated the left-out scene DeepMeaning map with a scene fixation density map that summarized where participants looked in that scene (Figure 3a, indoor mean R_{cv} = 0.63, bootstrap 95%CI [0.61, 0.65]; outdoor mean R_{cv} = 0.58, bootstrap 95%CI [0.53, 0.62]) and directly compared this to the correlation observed between human meaning maps and scene fixation density maps (Figure 3b, indoor mean R = 0.61, bootstrap 95%CI [0.59, 0.63]; outdoor mean R = 0.56, bootstrap 95%CI [0.52, 0.60]). Overall, DeepMeaning accounted for attention as well as human meaning maps for both indoor (t(98) = −1.14, p = 0.26, t-distribution 95%CI [−0.05, 0.01]) and outdoor scenes (t(98) = −0.76, p = 0.45, t-distribution 95%CI [−0.08, 0.04]). Moreover, there was a strong correlation (R = 0.75; Pearson correlation test, R(98) = 0.75, p < 0.001, 95%CI [0.64, 0.82], Figure 3c) of the scene-by-scene attention correlations for DeepMeaning and human meaning, indicating that DeepMeaning and human meaning maps also predicted attention very similarly for a given scene. Next, we tested the generalization performance of DeepMeaning by using an ensemble indoor and outdoor model (average linear weights and intercept across leave-one-scene-out folds) to estimate meaning in 100 indoor and 100 outdoor scenes from a second eye movement dataset with a smaller number of observers (CAT2000, Borji & Itti, 2015, Figure 3d). Again, we found that DeepMeaning maps were strongly associated with attention for both indoor (t(99) = −9459.20, p < .001, t-distribution 95%CI [0.50, 0.54]) and outdoor scenes (t(99) = −7558.35, p < .001, t-distribution 95%CI [0.42, 0.47]).

Having established that DeepMeaning accurately estimates local scene meaning and that DeepMeaning maps are strongly associated with where people look, we then tested whether DeepMeaning could detect the removal of local semantic information by applying a diffeomorphic transformation (Hayes & Henderson, 2022; Stojanoski & Cusack, 2014). The diffeomorphic transformation (Figures 4a and 4d) preserves the basic perceptual properties of the scene region while degrading its semantic content. Previously, we have shown that human meaning maps passed this semantic validity test, while three state-of-the-art deep saliency models failed (Hayes & Henderson, 2022). Therefore, for DeepMeaning to count as an automated method for estimating local scene meaning, DeepMeaning must also be able to pass this semantic validity test. To perform the diffeomorph test, we compared DeepMeaning's scene prediction for both the original scene and diffeomorphed scene using the same leave-one-scene-out cross-validation procedure as before (Figure 2a). We then compared the mean rating value for the critical region (original and diffeomorph) using a paired samples t-test. As can be seen (Figures 4b and 4c), DeepMeaning showed a large decrease in estimated meaning for the diffeomorphed region relative to the original unaltered scene region ($t(39) = 22.43$, $p < .001$, t-distribution 95%CI [0.68, 0.81], $d = 2.94$). This result demonstrates that DeepMeaning, like human raters, accurately detects decreases in local semantic content.

Semantic Interpretation

Finally, and perhaps most importantly, we evaluated whether the estimates DeepMeaning made were interpretable. In the first analysis, we show how the text generation ability of CoCa can be used to interpret our diffeomorph results as a simple proof-of-concept. While our second analysis leverages the full power of a multimodal shared vision-language space to define concrete semantic dimensions in the shared feature space (e.g., object density, interaction potential) and how these relate to DeepMeaning and human meaning estimates in the indoor scenes from our internal eye tracking dataset.

Given the Contrastive Captioner (CoCa) multimodal backbone of DeepMeaning, we can decode a local scene region into a text caption, providing human-interpretable insight into the model's representation of a given scene region. As a proof-of-concept of this ability, we decoded CoCa's representation for both the original and diffeomorphed scene patches into captions (e.g., Figure 4d) to understand why the DeepMeaning rating drops in the diffeomorphed region relative to the original in each scene. In all 40 original scene regions, semantic content was extracted (e.g., 'a shelf with many jars of food on it') with a caption accuracy of 92.5% (37/40), while producing semantically vacuous output for almost all (37/40) of the diffeomorphed image patches (e.g., 'a circular image of some sort with different colors'), indicating the model struggled to extract any semantic content from the diffeomorphed scene regions. A closer examination of the number of total objects that appear in each caption showed the original captions contained 103 objects compared to 16 objects in the diffeomorphed captions. This 84.5% drop in the number of objects represented provides a clear explanation for the large 2.94 standard deviation drop in the DeepMeaning ratings we observed when a region was diffeomorphed: when the amount of semantic content represented plummets, so does the DeepMeaning rating.

In addition to caption generation, we also used CoCa's shared vision-language embedding space to interpret DeepMeaning and human meaning ratings by quantifying the role of specific semantic directions we defined using contrastive prompts. We tested four semantic dimensions of local meaning using this semantic projection analysis: object density, interaction potential, figure-ground organization, and local contextual consistency (Figure 5 and Table 1). In this approach, for each dimension, we created simple contrastive prompts that defined a semantic direction in the CoCa embedding space (Figures 5a, d, g, j). Projecting each patch onto these semantic directions produced semantic scores for each scene patch that we compared to DeepMeaning estimates, human meaning ratings, and attention (Table 1). The scatter plots (Figures 5b, e, h, k) visualize the relationship between DeepMeaning ratings, human meaning ratings, with darker blue points indicating higher semantic projection scores as DeepMeaning and human meaning ratings increase. Together these semantic dimensions explained substantial total variance in DeepMeaning estimates ($R^2 = 0.56$), human meaning ratings ($R^2 = 0.47$), and attention ($R^2 = 0.29$). While the dominance analysis reveals that overall each semantic dimension explained about a quarter of the total variance (Table 1). This suggests that each of these semantic dimensions contribute approximately equally to local meaning.

Visual inspection of representative patches further validated our semantic projection analysis. The object density dimension showed a clear progression: high-density patches contained many objects, median patches showed moderate object counts, and low-density patches featured few objects (Figure 5c). The interaction dimension similarly ranged from highly manipulable objects (utensils, tools) through moderately interactive elements (handrails, chairs, bookcases) to static, non-interactive elements (Figure 5f). The figure-ground example patches demonstrated a continuum from near to distant views (Figure 5i), while the context dimension (Figure 5l) ranged from objects in typical settings to those appearing disconnected from their environment (such as a seemingly floating bucket). Together, the caption and semantic projection analyses demonstrate how a shared vision-language representational space can serve as a human-interpretable bridge between vision and language, providing a powerful tool for testing cognitive theory about semantic guidance in real-world scenes.

GENERAL DISCUSSION

Semantic representations are central to unraveling the interplay between scene understanding and visual attention. Previous work has approached this problem by either measuring direct human behaviors (i.e., semantic ratings of images and eye movement behavior relative to semantic feature manipulations) or by estimating human semantic representations based on regularities in large text corpora. Both approaches are useful, but they leave a representational gap that makes it difficult to determine the precise mapping between visual input and semantic knowledge, either because they are filtered through the human brain or because they are only based on vision or language without a mapping to the other.

By bridging vision and language representations, the present study achieves two goals. First, it offers an image-computable method for local scene meaning estimation that transfers to visual attention. This reduces the effort required to use tools like meaning maps because human ratings only need to be collected once and then a model can be trained that provides good estimates. Second, and perhaps more importantly, it provides a means to interpret the semantic representations that underlie the relationship between scene meaning and attention. Model interpretation is a major advantage for vision-language transformers over vision-only convolutional or transformer models that are difficult to interpret. More broadly, the current study serves as another piece of evidence that multimodal transformers like CoCa can serve as foundational vision-language models for downstream tasks (Yu et al., 2022).

We stress here that DeepMeaning is not to be taken as a global model of scene attention (Hayes & Henderson, 2022). Rather, it is a tool for testing hypotheses about the role of semantic features in attention. In this sense, DeepMeaning offers a fundamentally different approach than global fixation models (e.g., deep saliency models like DeepGaze, Kümmerer et al., 2016) where the primary goal is to capture the maximum amount of variance possible in attention by training deep neural network models directly on fixation data (Henderson et al., 2021). In our view, the model benchmarking approach using deep neural networks (e.g., CNNs or transformers) is not particularly conducive to informing theories of attention and scene understanding, because it does not tell us which specific features contribute to a model's success (Bowers et al., 2022; Hayes & Henderson, 2021a). Importantly, unlike other models, DeepMeaning never sees fixation data, and this is precisely why it is theoretically interesting that it transfers to predict fixation behavior. Meaning maps test hypotheses concerning the relationship between scene semantics and attention by isolating semantic features like local recognizability/informativeness (Henderson & Hayes, 2017), graspability (Rehrig et al., 2020), and interactability (Rehrig et al., 2022) to understand their unique roles, not by out-predicting other models on a benchmark dataset. We believe DeepMeaning provides an important new image-computable method and interpretive framework for testing specific hypotheses about the role of semantics in scene understanding and the guidance of attention.

Our results demonstrate how vision-language transformers can decompose local meaning into interpretable semantic components that transfer to attention. We tested four semantic dimensions: object density, interaction potential, figure-ground organization, and local context consistency. These four components alone explained over half the explainable variance in DeepMeaning's estimates, human meaning ratings, and attention. The strong relationship between these semantic dimensions and attention suggests that viewers' gaze is guided by readily identifiable object properties: how many objects are present, their affordances for interaction, their spatial position within the scene relative to the observer, and their contextual relationships with the immediately surrounding objects. Local meaning has been established as a robust predictor of attention across a wide range of viewing tasks (Henderson et al., 2019), including tasks where local meaning is task-irrelevant (Hayes & Henderson, 2019; Peacock et al., 2019). The present work provides concrete insights into how semantic features shape attention allocation during scene viewing by decomposing local meaning using a vision-language computational framework.

Importantly, the usefulness of a vision-language transformer framework extends beyond meaning maps to offer a general tool for investigating the role of semantics on visually guided behaviors. For any visual task where semantics may guide attention or behavior (e.g., accuracy or reaction time), researchers can generate 'semantic prompt maps' by scoring local image regions along theoretically-driven or hypothesis-driven semantic dimensions using contrastive prompts just like we have done here. For example, rather than object interactability, one could map regions based on their task relevance or emotional valence. These semantic prompt maps could be used to both test existing theories and to generate new hypotheses about how specific semantic properties influence aspects of visually-guided behaviors. This suggests that integrating across vision and language, rather than remaining siloed in a single domain, represents a promising direction for advancing our understanding of semantics in visual cognition.

While very promising, the semantic prompt map approach does require the user to define contrastive prompts to target specific semantic dimensions. The effectiveness of contrastive prompts at capturing specific semantic dimensions can vary depending on the quality of the prompts themselves. While we did not directly validate our semantic prompts with direct human ratings (e.g., Grand et al., 2022), our approach benefits from an indirect validation through the correlations between our semantic model dimensions and human meaning ratings. When applying this methodology to novel semantic dimensions without comparable human rating data, researchers may benefit from additional validation steps to ensure the contrastive prompts are capturing the intended semantic dimensions. This consideration does not diminish the overall utility of the semantic prompt map approach but rather highlights an area for thoughtful implementation when adapting this approach to different aspects of semantic guidance.

In summary, we used a state-of-the-art vision-language transformer trained on billions of image-text pairs to investigate how joint representations learned from vision and language can predict what scene regions people find meaningful and consequently where they look. We demonstrated that this computational framework successfully recovers human meaning ratings, transfers as a strong predictor of scene attention, detects local changes in semantic content like human raters, and provides a direct route to interpreting the components of local meaning that guide attention. The ability to offer image-computable local scene meaning estimation with a concrete pathway for model interpretation has tremendous potential for advancing our understanding of how semantic representations produce rapid scene understanding and help guide attention, with direct implications for cognitive science, computer vision, linguistics, robotics, and artificial intelligence.

ACKNOWLEDGMENTS

This research was supported by the National Eye Institute of the National Institutes of Health, under award number R01EY027792. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

TRH and JMH conceived the study, TRH conducted the experiments and analyzed the data. TRH drafted the manuscript and JMH revised the manuscript.

DATA AVAILABILITY STATEMENT

All Python code and data required to reproduce the results from this paper are available here: https://OSF.IO/HCNFX.

REFERENCES

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70. https://doi.org/10.1037/h0036799, PubMed: 4424680

Azen, R., & Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129, PubMed: 12924811

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77–80. https://doi.org/10.1126/science.177.4043.77, PubMed: 5041781

Borji, A., & Itti, L. (2015). CAT2000: A large scale fixation dataset for boosting saliency research. In *CVPR 2015 workshop on "Future of Datasets". arXiv preprint arXiv:1505.03581*.

Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. https://doi.org/10.1017/S0140525X22002813, PubMed: 36453586

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. Retrieved from https://arxiv.org/abs/2104.14294.

Clarke, A., & Tyler, L. K. (2015). Understanding what we see: How we derive meaning from vision. *Trends in Cognitive Sciences*, *19*(11), 677–687. https://doi.org/10.1016/j.tics.2015.08.008, PubMed: 26440124

Cronin, D. A., Hall, E. H., Goold, J. E., Hayes, T. R., & Henderson, J. M. (2020). Eye movements in real-world scene photographs: General characteristics and effects of viewing task. *Frontiers in Psychology*, *10*, 2915. https://doi.org/10.3389/fpsyg.2019.02915, PubMed: 32010016

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*. https://doi.org/10.48550/arXiv.2010.11929

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8, PubMed: 35422527

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372, PubMed: 31505121

Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin and Review*, *26*(5), 1683–1689. https://doi.org/10.3758/s13423-019-01642-5, PubMed: 31342407

Hayes, T. R., & Henderson, J. M. (2021a). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific Reports*, *11*(1), 18434. https://doi.org/10.1038/s41598-021-97879-z, PubMed: 34531484

Hayes, T. R., & Henderson, J. M. (2021b). Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, *32*(8), 1262–1270. https://doi.org/10.1177/0956797621994768, PubMed: 34252325

Hayes, T. R., & Henderson, J. M. (2022). Meaning maps detect the removal of local semantic scene content but deep saliency models do not. *Attention, Perception, & Psychophysics*, *84*(3), 647–654. https://doi.org/10.3758/s13414-021-02395-x, PubMed: 35138579

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. Retrieved from https://arxiv.org/abs/2111.06377.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006, PubMed: 14585447

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*, 219–222. https://doi.org/10.1111/j.1467-8721.2007.00507.x

Henderson, J. M. (2011). Eye movements and scene perception. In I. S. P. Liversedge, D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 594–606). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0033

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747. https://doi.org/10.1038/s41562-017-0208-0, PubMed: 31024101

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10. https://doi.org/10.1167/18.6.10, PubMed: 30029216

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), 19. https://doi.org/10.3390/vision3020019, PubMed: 31735820

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kummerer, Wallis, Bethge & Teufel (2021). *Cognition*, *214*, 104742. https://doi.org/10.1016/j.cognition.2021.104742, PubMed: 33892912

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243, PubMed: 10074679

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2007). Cortical activation to indoor versus outdoor scenes: An fMRI study. *Experimental Brain Research*, *179*(1), 75–84. https://doi.org/10.1007/s00221-006-0766-2, PubMed: 17123070

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., … Schmidt, L. (2021). *OpenCLIP*. Zenodo. https://doi.org/10.5281/zenodo.5143773

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *CoRR, abs/1610.01563*. https://doi.org/10.48550/arXiv.1610.01563

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3565–3565. https://doi.org/10.1016/S0042-6989(01)00102-X, PubMed: 11718795

Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common objects in context. *CoRR, abs/1405.0312*. https://doi.org/10.48550/arXiv.1405.0312

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565–572. https://doi.org/10.1037/0096-1523.4.4.565, PubMed: 722248

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*(11), 547–552. https://doi.org/10.3758/BF03210264

Murphy, G. (2004). *The big book of concepts*. MIT Press.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*(1), 20–34. https://doi.org/10.3758/s13414-018-1607-7, PubMed: 30353498

Potter, M. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966. https://doi.org/10.1126/science.1145183, PubMed: 1145183

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rodgers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150, PubMed: 27881854

Rehrig, G., Barker, M., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2022). Look at what I can do: Object affordances guide visual attention while speakers describe potential actions. *Attention, Perception, & Psychophysics*, *84*(5), 1583–1610. https://doi.org/10.3758/s13414-022-02467-6, PubMed: 35484443

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J., & Ferreira, F. (2020). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1659–1681. https://doi.org/10.1037/xlm0000837, PubMed: 32271065

Reilly, J., Shain, C., Borghesani, V., Kuhnke, P., Vigliocco, G., Peelle, J. E., Mahon, B. Z., Buxbaum, L., Majid, A., Brysbaert, M., Borghi, A. M., Deyne, S. D., Dove, G., Papeo, L., Pexman, P. M., Poeppel, D., Lupyan, G., Boggio, P., Hickok, G., … Vinson, D. (2025). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review*, *32*(1), 243–280. https://doi.org/10.3758/s13423-024-02556-7, PubMed: 39231896

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth conference on neural information processing systems datasets and benchmarks track* Retrieved from https://openreview.net/forum?id=M3Y74vmsMcY.

SR Research. (2010). *EyeLink 1000 user's manual, version 1.5.2*. Mississauga, ON: SR Research Ltd.

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12), 6. https://doi.org/10.1167/14.12.6, PubMed: 25301014

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766, PubMed: 17014302

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009, PubMed: 31051430

Williams, C. C., & Castelhano, M. S. (2019). The Changing Landscape: High-level Influences on Eye Movement Guidance in Scenes. *Vision*, *3*(3), 33. https://doi.org/10.3390/vision3030033, PubMed: 31735834

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 0058. https://doi.org/10.1038/s41562-017-0058, PubMed: 36711068

Wu, C.-C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 54. https://doi.org/10.3389/fpsyg.2014.00054, PubMed: 24567724

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum. https://doi.org/10.1007/978-1-4899-5379-7

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). CoCa: Contrastive captioners are image-text foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2205.01917