Looking for Semantic Similarity: What a Vector Space Model of Semantics Can Tell Us About Attention in Real-world Scenes (Supplementary Materials)

Subject- and Scene-level variation in Concept and Center Proximity Map Values

Greater subject-level variation was observed in the fixated center proximity values than concept map values (Fig. S1a). The opposite pattern was observed in the scene-to-scene variability (Fig. S1b). That is, the fixated concept map values showed greater variability scene-to-scene than the center proximity values. These findings highlight the importance of accounting for the random effects of subject and scene in our statistical model.

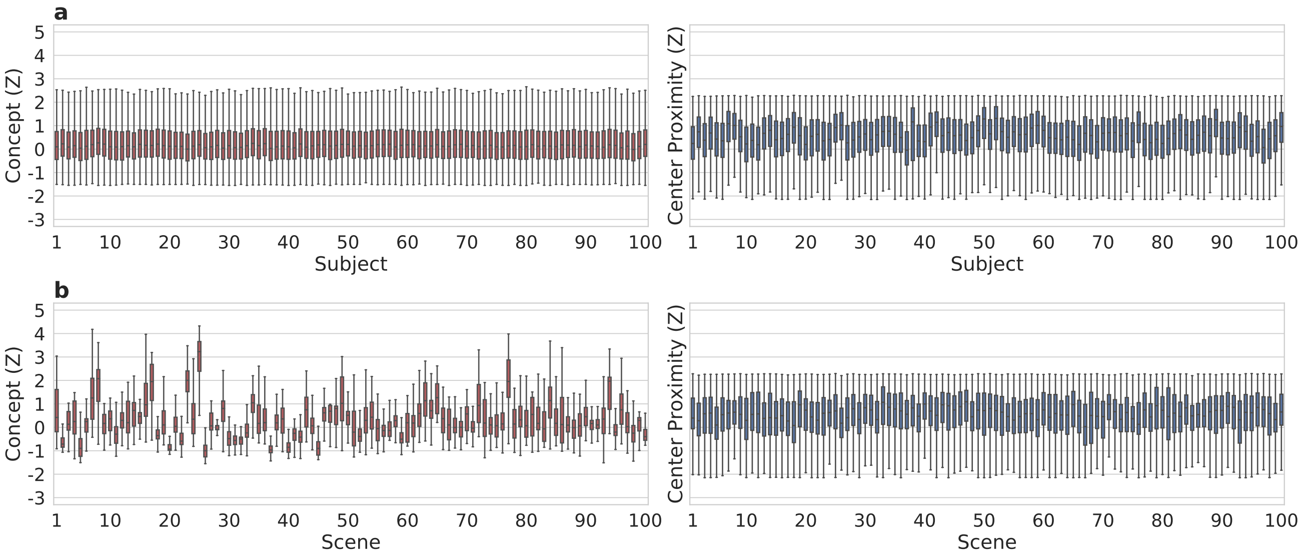

Fig. S1 *Box plots illustrating the subject-level (a) and scene-level (b) variation in the fixated concept and center proximity map values.* The box plot represents one subject or scene. The extent of the colored box indicates the 1st and 3rd quartiles, the vertical black whiskers show the minimum and maximum values, and the horizontal black line indicates the median.

Fixated/Non-fixated Distinction

In order to use a logistic general linear mixed effects modeling framework to model the relationship between scene features and eye movements requires both fixated and non-fixated scene locations for each subject and scene (Nuthmann, Einhäuser, & Schütz, 2017). In a previously proposed mixed effects approach by Nuthmann et al. (2017), they split each scene into an arbitrary 8x6 grid and then examined fixated and non-fixated grid scene features. This approach has two undesirable properties. First, using a grid creates a coarser fixation representation that is subject to boundary issues. That is, if two fixations fall near a grid boundary they can be assigned very different feature values despite falling on virtually identical locations in the scene. Second, as the number of fixations a subject makes in a scene grows, the likelihood that there will be any grid locations that were not fixated decreases. Our method of sampling a 3° window of the scene features around each region a subject fixated and did not fixate eliminates the boundary problem and is suitable for scenes with longer viewing durations (e.g., 12 seconds in our case).

GLME model diagnostics

We verified that the random effects in the GLME were normally distributed using quantile-quantile (Q-Q) plots. In a Q-Q plot, if the random effects are normally distributed, then the subject and scene random effects estimates should fall along the diagonal of the Q-Q plot. The Q-Q plots for both the subject and scene random effects indicated the assumption of normality was met (see Fig. S2a and S2b).

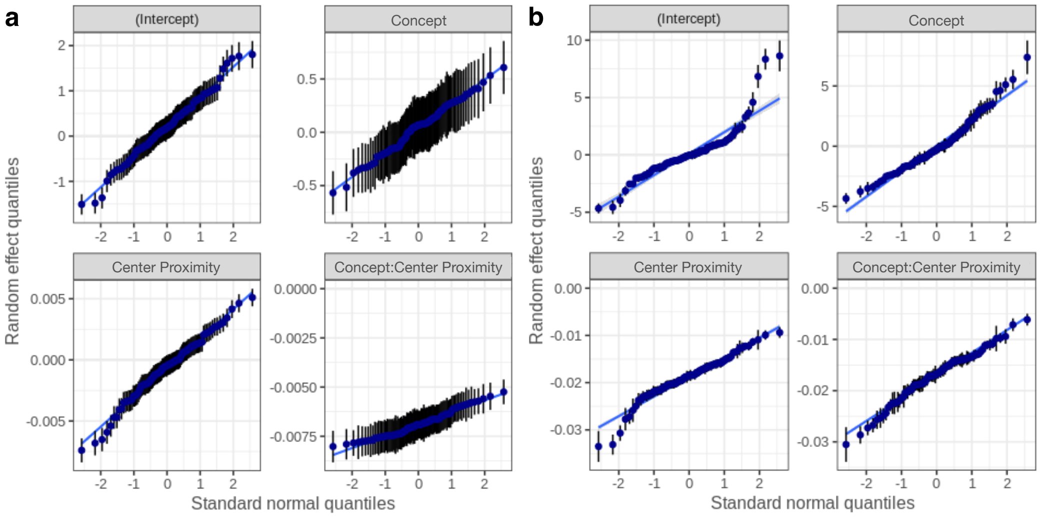

Fig. S2 *Quantile-quantile plots of the general linear mixed-effects model random effects* The quantile-quantile plots of the subject (a) and scene (b) random effects estimates for the intercept, concept, center proximity, and concept x center proximity interaction terms.

Differences in physical salience do not account for the semantic similarity effect

To test the possibility that the semantic similarity effect we observed was driven by differences in physical salience, we fit an additional GLME model to our data that included low-level saliency in the place of center proximity. Specifically, we fit a GLME logit model in which whether a scene region was fixated or not served as the dependent variable while the concept map value, low-level saliency values (as indexed by the Graph-based visual saliency model), and their interaction were fixed effects. Subject and scene again were treated as full random effects (i.e., slope and intercept). Center proximity was excluded to allow for the model to converge and avoid the interpretive difficulties of 3-way interactions among terms. The GLME model results indicated no significant interaction between low-level saliency map and the concept maps (see Fig. S3). This finding suggests that the relationship between the concept maps and fixations is not accounted for by differences in low-level features captured by image salience.

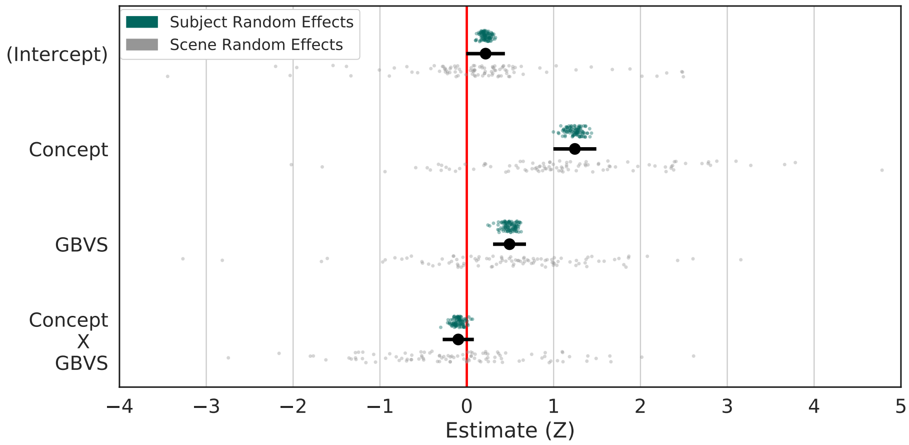

Fig. S3 *Fixation location general linear mixed-effects model (GLME) results.* In this supplementary GLME, whether a scene region was fixated or not served as the dependent variable while the concept map value, physical salience value, and their interaction were fixed effects. The black dots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (green dots) and scene (grey dots) were both accounted for in the model as random effects (intercept and slope).

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	z-value	p	by-subject	by-scene
Intercept	0.22	[-0.01 0.44]	0.12	1.89	0.06	1.16	1.12
Concept	1.25	[0.99 1.50]	0.13	9.63	< 0.001***	0.09	1.31
GBVS	0.49	[0.30 0.68]	0.10	5.09	< 0.001***	0.08	0.98
Concept x GBVS	-0.10	[-0.28 0.08]	0.09	-1.09	< 0.28	0.07	0.91

Supplementary Tab. 1 *Fixation location general linear mixed-effects model physical salience results.* Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $z-$values, and p-values (p) for each fixed effect and standard deviations (SD) for the random effects of subject and scene.

Computing object shape similarity

In an exploratory analysis, we also examined how high-level features like object shape were related to the object semantics from ConceptNet Numberbatch. To extract each object's shape contour, each object was cropped from the scene using a tight fitting bounding box, centered, and then scaled to a common array size that preserved aspect ratio. Together, these procedures produced a binary object shape contour for each object (Fig. S4a). Then, the pairwise shape similarity between the object contours were computed as their inverse Hamming distance (Fig. S4b). The inverse Hamming distance varies from 0 to 1 and indicates the proportion of agreeing pixels in the shape masks, with an inverse Hamming distance of 1 indicating identical shapes.

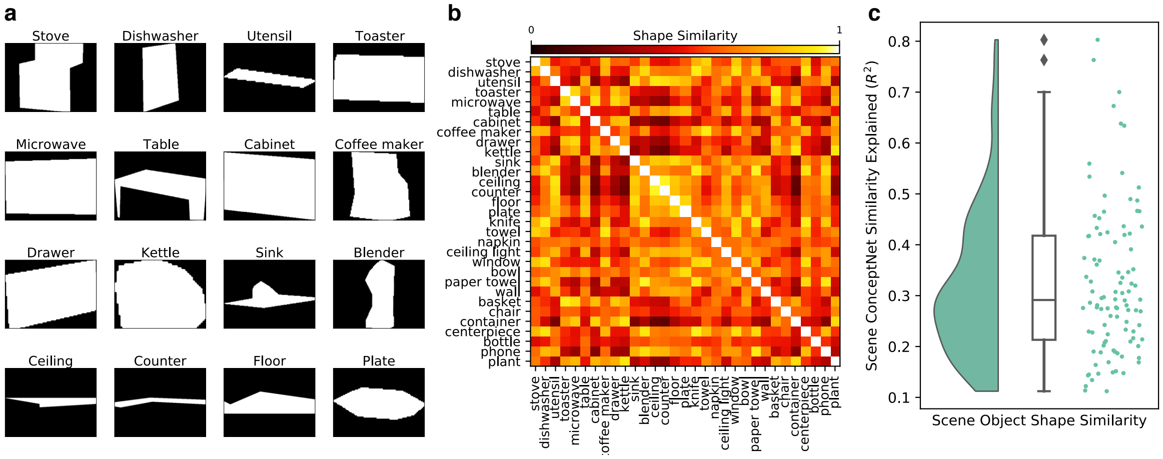

Fig. S4 *Object shape similarity analysis.* For the shape similarity analysis object contour masks (a) were first extracted from the scene. Then the proportion of shared pixels (i.e., inverse Hamming Distance) was used to compute pairwise object similarity (b). Shape similarity accounted for approximately a third of the variance (R^2) on average in the ConceptNet similarity matrices (c) suggesting object shape is related to the stored similarity structures in ConceptNet.

Shape similarity analysis

In the exploratory analysis, we tested whether the pairwise ConceptNet object similarity matrices (Fig. 1b) were related to the corresponding object shape similarity matrices (Fig. S4b) by simply correlating the two pairwise matrices together for each scene. The results (Fig. S4c) showed that object shape similarity explained about a third of the variance (R^2) in the object ConceptNet similarity values (M=0.32, SD=0.15; $t(99) = 22.09, p < .001$, 95% CI $[0.29, 0.35]$).

One interesting question that arises from our finding that object coherence guides attention is how do we actually use stored object similarity structures to guide our attention to semantically similar objects? All objects could be preattentively tagged, but we (Henderson & Hayes, 2017,2018) believe this is highly unlikely given previous findings from the scene gist (Oliva & Torralba,2006) and visual search (Wolfe, 2018) literature. The exploratory object shape analysis offers an intriguing potential answer that highly diagnostic visual features like object shape could serve as part of the mechanism by which semantic similarity representations facilitate attentional guidance. This hypothesis would be consistent with previous literature showing that conceptually similar objects also tend to be more visually similar (Snodgrass & Vanderwart, 1980; Potter & Faulconer,1975; Duncan & Humphreys, 1989).