

ARTICLE

How do visual and conceptual factors predict the object content in typical scene drawings?

Gongting Wang^{1,2}  | Ilker Duyamaz² | Matthew J. Foxwell³ |
 Micha Engeser^{2,4} | David Pitcher³ | Radoslaw M. Cichy¹ |
 Daniel Kaiser^{2,4,5}

¹Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

²Department of Mathematics and Computer Science, Physics, Geography, Justus-Liebig-Universität Gießen, Gießen, Germany

³Department of Psychology, University of York, York, UK

⁴Center for Mind, Brain and Behavior (CMBB), Justus-Liebig-Universität Gießen, Philipps-Universität Marburg and Technische Universität Darmstadt, Marburg, Germany

⁵Cluster of Excellence "The Adaptive Mind", Justus-Liebig-Universität Gießen, Philipps-Universität Marburg and Technische Universität Darmstadt, Gießen, Germany

Correspondence

Gongting Wang, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany.

Email: generalwgt@gmail.com

Funding information

Deutsche Forschungsgemeinschaft; European Research Council

Abstract

Imagine you draw a typical bedroom, your choice of objects is likely to depend on visual occurrence statistics (i.e. the objects present in previously encountered bedrooms) and semantic relations between objects and scenes (i.e. the semantic relationship between the bedroom and its constituent objects). To investigate how these two factors contribute to the composition of typical scene drawings, we analysed 1192 drawings of six indoor scene categories, obtained from 303 participants. For each object featured in the drawings, we estimated its visual occurrence frequency from the ADE20K dataset, and its semantic relatedness to the scene concept from a word2vec model. Across all scenes of a given category, generalized linear models revealed that visual and conceptual factors both predicted the likelihood of an object featuring in the scene drawings, with a combined model outperforming both single-factor models. We further computed the visual and semantic specificity of objects for a given scene, that is, how diagnostic an object is for the scene. Object specificity offered only weak predictive power when predicting the selection of objects, yet even infrequently drawn objects remained diagnostic of their scenes. Taken together, we show that visual and conceptual factors jointly shape the composition of typical scene drawings.

KEYWORDS

line drawings, real-world statistics, semantic scene composition, visual typicality, world models

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *British Journal of Psychology* published by John Wiley & Sons Ltd on behalf of The British Psychological Society.

BACKGROUND

Drawing is a fundamental form of human communication. Humans have employed drawings to communicate ideas, express emotions and share experiences for at least 40,000 years (Aubert et al., 2014; Hoffmann et al., 2018). In modern society, drawings continue to serve diverse functions. For instance, designers use drawings to visualize ideas (Goldschmidt, 2014), artists use drawings to externalize imagination (Fish & Scrivener, 1990), clinicians use drawings to diagnose and classify neurological disorders (Agrell & Dehlin, 1998; Wechsler, 2009). Far from being a merely preparatory or decorative act, drawing is a primary mode of visual thinking that renders thought visible across art, architecture, design, engineering and the sciences (Petherbridge, 2010). Contrasting this prevalence of drawings, we know relatively little about how people decide which elements to include when composing a drawing. This question is particularly evident when drawings of complex scenes are considered: which objects do people draw to convey a certain scene, say a living room? A better understanding of the composition of such drawings may shed a new light on how people translate perceptual and memory representations into visible outputs.

Such insights are also crucial for researchers in cognitive science and neuroscience that use drawing as a tool to characterize the nature of internal representations (Engeser et al., 2025; Fan et al., 2023; Roberts & Wammes, 2021). For instance, in development research, drawings are used to investigate the developmental trajectory of visual object representations (Karmiloff-Smith, 1990; Long et al., 2019, 2024). In memory research, successes and failures in memory can be captured by how well human drawings during recall align with or deviate from the studied materials (Bainbridge et al., 2019; Bainbridge & Baker, 2020; Fan et al., 2023; Metzger, 1936). Further, in perception research, drawings provide descriptions of participants' world models, which can in turn be used to predict perception (Engeser et al., 2025; Morgan et al., 2019; Wang et al., 2024, 2025). A better understanding of how people compose their drawings could inform the potential and limitations of studies using drawing as a methodological tool.

To understand how people compose drawings of natural scenes, we analysed more than 1000 drawings of six scene categories from more than 300 participants. All participants were asked to draw a typical instance of a scene category (e.g. a typical living room), after briefly thinking about the scene contents and then drawing within a relatively liberal time constraint (Wang et al., 2024, 2025). The full set of scene drawings, the *Room Drawings Dataset*, is released alongside this publication (see Materials and Methods), providing a rich benchmark dataset of evaluating diverse aspects of drawing composition in future studies.

In the current study, we used the dataset to ask how well the object composition in drawings from each scene category could be predicted by two complementary factors: (i) *visual occurrence statistics*, that is, the frequency with which the individual objects are encountered in a given scene category in the real-world, and (ii) *semantic similarity*, that is, the semantic relatedness between the scene concept and the individual object concepts.

First, people's decisions about which objects are drawn for a given category should predict on how frequently the objects are commonly found in scenes of that category. In bathrooms, we much more often see a sink than a table, so we should draw sinks more often than tables. Most scenes are reliably associated with such prominent scene- and space-defining objects (Bar, 2004; Bar & Aminoff, 2003; Oliva & Torralba, 2007; Võ et al., 2019). Here, we quantified such *visual occurrence statistics* by assessing how frequently individual objects appear within images real-world scene categories. Specifically, we used the annotated ADE20K scene database (Zhou et al., 2016, 2017) and computed for a set of scene categories how often individual objects featured in images of that category (e.g. how often is a sink found in images of bathrooms).

Second, the object composition of a drawing may not just be shaped by visual occurrence statistics. It may also depend on information stored in conceptual representations, which do not necessarily mirror unfiltered visual experience. During drawing, people need to recall relevant objects from long-term memory. Such long-term memory for real-world objects can be stored in structured conceptual spaces

(Brady et al., 2008; Konkle et al., 2010), and retrieval from such conceptual spaces may depend on the semantic similarity between a scene concept and the candidate object concepts. Here, we quantified such *semantic similarity* by assessing the similarity of object and scene concepts in a large language model (word2vec; Mikolov et al., 2013). Specifically, we computed the cosine similarity between a set of scene category concepts and individual object concepts in this model (e.g. how similar is the concept ‘sink’ to the concept ‘bathroom’).

These two predictors enabled us to test how visual occurrence statistics and semantic relatedness contribute to the object composition in a set of more than 1000 scene drawings spanning six categories.

MATERIALS AND METHODS

Participants

A total of 303 participants (26.14 ± 4.78 years $\pm$ SD, 100/201/2 male/female/other) provided scene drawings. Of these, 101 participants (25.20 ± 4.07 years $\pm$ SD, 24/77 male/female) were tested online in the United Kingdom (recruited at the University of York), and 202 participants (26.62 ± 5.04 years, 76/124/2 male/female/other) were tested in a laboratory setting in Germany (recruited at Freie Universität Berlin and Justus-Liebig-Universität Giessen). Procedures were approved by the ethics committees of the Department of Psychology, University of York, the Department of Education and Psychology, Freie Universität Berlin, and the ethics committee of the Justus-Liebig-Universität Gießen, respectively and adhered to the Declaration of Helsinki.

Drawing sessions

The drawings used here were produced in drawing sessions across multiple experiments, including two published studies (Wang et al., 2024, 2025) and other still unpublished work. Drawing sessions were conducted either online or in-person in a laboratory setting. Online sessions were conducted via Skype. Here, participants provided their drawings using a pencil, eraser and ruler on A4 paper. Drawings were produced within a standardized perspective grid to support depth depiction and to match viewpoints across scenes. In the online session, participants drew or printed grids themselves on A4 paper and consisted of a large central rectangle (7.1 cm by 16.5 cm) and four diagonal lines going from each corner of the rectangle to the corner of the page. The rectangle was placed 8.5 cm from the bottom of the page and 5.4 cm from the top of the page. For each drawing, participants had 1 min to plan and 3 min 30 s to complete the drawing. Participants drew typical versions of three scene categories (bedroom, kitchen, living room). In the lab-based drawing sessions, participants provided their drawings using an Apple Pencil on an Apple iPad with the Sketchbook App. In the lab-based session, a standardized perspective grid similar to the one used in the online session was provided for each drawing. Specifically, the full drawing display (19.5 cm by 26.1 cm) consisted of a large central rectangle (8.5 cm by 13.5 cm) and four diagonal lines from each corner of the rectangle to the corner of the page. The rectangle was set with the bottom length 8.5 cm from the bottom of the page and top length 2.7 cm from the top of the page. Each participant was given 30 s to plan and 4 min to complete each drawing. A group of participants ($N=85$) drew six typical scene categories (bathroom, bedroom, café, kitchen, living room, office), while another group ($N=115$) drew four categories (bathroom, bedroom, kitchen, living room). In all sessions, participants were instructed to draw the most typical representation of each room (Figure 1), rather than their own rooms or an aesthetically appealing version of the room. Participants were instructed to draw into a pre-defined perspective grid, which they drew themselves according to the experimenter's instructions (online sessions) or which was present on the iPad screen from the outset (lab-based sessions). In the lab-based sessions, participants were instructed to use only the Apple Pencil and the eraser tool in Sketchbook, replicating the affordances of pencil and paper

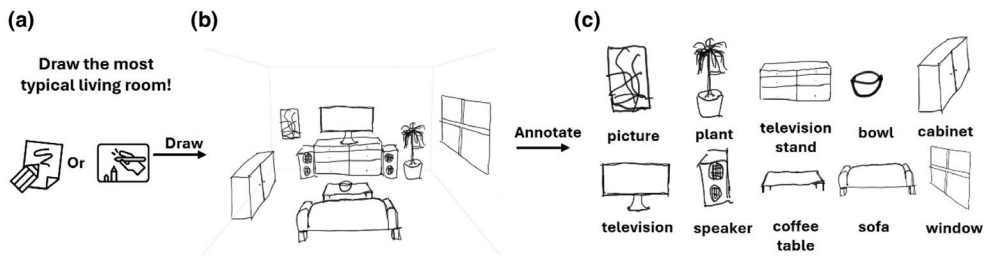


FIGURE 1 Overview of the drawing sessions and annotation procedure. (a) Participants draw typical versions of indoor scene categories on paper or an iPad. (b) An example drawing of a living room. Note that a common perspective was enforced by a perspective grid shown to the participants from the outset. (c) We manually annotated all individual objects in the drawings. Each object instance was only counted once (i.e. objects were coded as present or absent).

drawing. Object manipulation features (e.g. moving, rescaling) were not permitted. Further details on the drawing sessions are available in our previous studies (Wang et al., 2024, 2025).

Object annotation

At the end of each drawing session, the experimenter went through each drawing with the participant to verify the identity of every depicted object. For any object whose identity was visually ambiguous, the participant was asked to provide a written label, which was then used as the annotation. For each drawing, all depicted objects were manually annotated. Overall, typical bathroom drawings consisted of an average of 6.3 different objects ($SD=1.6$, range=[4, 12]; multiple instances of the same object were not counted), with the three most frequent objects sink (94.7%), toilet (90%) and shower (83.5%). Bedrooms consisted of an average of 8.3 objects ($SD=2.1$, range=[4, 13]), with the three most frequent objects bed (98.4%), pillow (83.2%) and window (74%). Cafés consisted of an average of 7.4 objects ($SD=3.1$, range=[3, 19]), with the three most frequent objects table (98.0%), chair (95.0%) and counter (76.2%). Kitchens consisted of an average of 8.7 objects ($SD=3.4$, range=[4, 17]), with the three most frequent objects cupboard (89%), stove (87.7%) and sink (79%). Living rooms consisted of an average of 7.8 objects ($SD=2.5$, range=[4, 14]), with the three most frequent objects sofa (99.6%), television (87.6%) and television stand (70.9%). Offices consisted of an average of 7.3 objects ($SD=3.1$, range=[4, 15]), with the three most frequent objects table (99.0%), chair (95.9%) and window (74.2%). Full annotation details are provided in the [supplementary materials](#). From these annotations, we computed the occurrence frequency of each object within its respective scene category.

Quantifying visual occurrence statistics and semantic relatedness

To quantify visual occurrence statistics across the exemplars of each scene category, we determined the occurrence frequencies of the annotated objects in each scene category by referencing the ADE20K dataset (Zhou et al., 2016, 2017). Specifically, we queried the database for each object annotated in the drawings and the corresponding scene category and calculated each object's frequency of occurrence across all the scenes for each category. To quantify semantic relatedness between object and scene concepts, we computed the similarity between the word representing each object annotated in the drawings (e.g. 'shower' in English, 'Dusche' in German) and the corresponding scene concepts (e.g. 'bathroom' in English, 'Badezimmer' in German) using a word2vec model. We used word vectors pre-trained on German for the German participants and on English for the UK participants, and supplied words in each group's native language. Both training resources came from the Common Crawl and Wikipedia corpora using fastText (Grave

et al., 2018). By calculating the cosine similarity between the object and scene concept, we quantified how strongly each object is semantically related to the scene category it appeared in.

Modelling object drawing frequencies

To examine the relationship between object drawing frequency, visual occurrence statistics and semantic relatedness, we fitted generalized linear models to predict object drawing frequencies using (i) occurrence statistics only, (ii) semantic relatedness only and (iii) both predictors together. As the dependent variable represented a bounded proportion (i.e. % of drawings that contained the object), we chose a Beta-binomial regression approach (i.e. a generalized linear model with a Beta function as the link function, Ferrari & Cribari-Neto, 2004). To assess the overall effect of visual experience and conceptual knowledge on object occurrence frequency across all categories, we fitted a generalized linear mixed-effects model that included category as a random effect as well as visual experience and/or conceptual knowledge as fixed effects. We then assessed whether a combined model better explained drawing frequencies better than occurrence statistics or semantic relatedness alone. To assess the stability of the results across categories, we further fitted the same model individually for every scene category.

We explored the composition of the drawings further by asking how the specificity of an object for a given scene category (e.g. a stove is highly specific for a kitchen, as it almost exclusively appears there, but a window is not) predicts drawing frequency. To quantify specificity, we calculated (i) the scene-specificity of an object in visual occurrence statistics, defined as the normalized difference between an object's frequency in its corresponding scene category and its average frequency across the other scene categories, (ii) the scene-specificity of an object in semantic relatedness, defined as the normalized difference between an object concept's cosine similarity to its corresponding scene concept and its average cosine similarity to the other scene concepts in the language model. Specificity was computed using the following formula:

$$SpecF/S(o|s) = \frac{F/S(o,s) - \overline{F/S(o,-s)}}{F/S(o,s) + \overline{F/S(o,-s)}}$$

F: object frequency, S: semantic relatedness, o: object, s: corresponding scene, -s: the other scenes.

Then, we fitted mixed-effects models to predict object frequency in drawings from visual and semantic specificity. We further asked whether infrequently drawn objects are still diagnostic for their respective scene category (and if so, to which extent) or whether they mainly constitute generic 'filler' objects that are equally appropriate for our range of categories (like windows or bins). To this end, we ordered all objects by their drawing frequency and binned them into six ranges: objects featured in 0%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–70% or 70%–100% of scenes. We varied bin sizes across the frequency range, as there were more objects that appeared relatively infrequently. We then assessed whether objects across frequency bins were consistently diagnostic for the scene category.

The ADE20K annotations analysis and the word2vec analysis were conducted in Python. All further statistical analyses were conducted in R.

Data, material and code availability

Data and code are accessible on the Open Science Framework (OSF), available at <https://osf.io/p2fa6/>. We further release all drawings, together with participant information and annotations, in the *Room Drawings Dataset*, available at: <https://osf.io/byu24/>.

RESULTS

We first visualized how often different objects were drawn for each scene category. We generated word clouds (Heimerl et al., 2014) for each category, in which font size reflects how frequently each object was drawn (Figure 2). These descriptive data show that each scene featured prominent and diagnostic objects that were drawn across many instances of the scene (e.g. a bed in the bedroom).

To evaluate how visual occurrence statistics and semantic relatedness contribute to the composition of scene drawings, we fitted three mixed-effects models that predicted object drawing frequency from either (i) only visual occurrence statistics, (ii) only semantic relatedness or (iii) both factors combined. Each model additionally featured category as a random effect. Model estimates and fit indices are reported in [Table 1](#) (‘Full Model’). A comparison of Akaike information criterion (AIC) and Bayesian information criteria (BIC) revealed that the combined model provided the best fit (AIC = -679.50, BIC = -659.98), with a conditional R^2 of 0.81. These results suggest that visual occurrence statistics are a good predictor of object drawing frequencies in scene drawings. Yet, semantic similarity between scene and object concepts explains additional variance in the drawing composition that is not captured by visual statistics.

Moreover, the random effect of category was significant ($p < 0.001$). We thus examined whether predictor contributions varied across scene categories; we then fitted separate generalized linear models with both predictors in each category (Figure 3). In all cases, both visual occurrence statistics and semantic relatedness remained significant, although their relative contributions varied across categories (Table 1). The generalized linear models also predicted object drawing frequency when trained on all but one category and tested on the remaining category (see Supplementary Information), suggesting an overall similar influence of visual occurrence statistics and semantic relatedness across categories. Finally, the complementary contribution of visual and semantic factors was corroborated by a simpler partial correlation analysis, where partialling out visual occurrence statistics still yielded significant correlations between semantic similarity and object drawing frequencies, and vice versa (see Supplementary Materials).

While the above analyses show that both visual and semantic factors contribute to whether objects are features in a drawing, it remains unclear whether participants select objects primarily based on absolute frequencies (i.e. how often does a chair appear in a living room) or based on relative frequencies

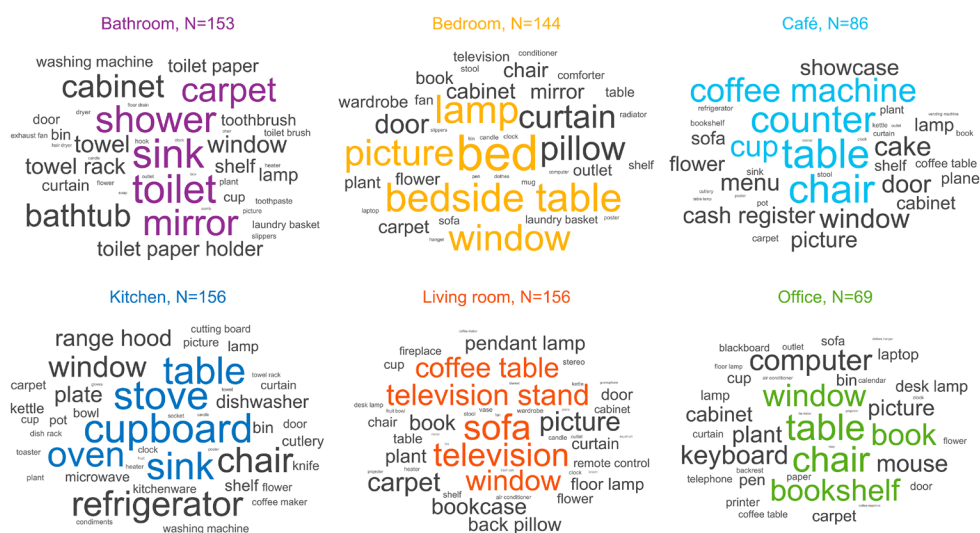


FIGURE 2 Word clouds representing the drawing frequency of objects across scene categories. The size of the words represents how frequently the object was featured in the drawings.

TABLE 1 Regression weights and goodness-of-fit statistics for the generalized linear models.

Scene category	Visual	Semantic	AIC	BIC	R ²
Full Model	4.78***	/	−645.65	−630.04	0.77
	/	6.59***	−436.85	−421.24	0.54
	4.33***	3.21***	−679.50	−659.98	0.81
Bathroom	4.39***	5.99***	−92.89	−86.12	0.67
Bedroom	4.56***	2.10*	−156.11	−146.39	0.64
Café	6.03***	10.37***	−56.86	−50.52	0.65
Kitchen	4.36***	2.44*	−138.78	−129.10	0.61
Living room	3.63***	3.30***	−158.63	−148.86	0.51
Office	6.69***	2.69*	−89.46	−82.91	0.74

* $p < 0.05$. *** $p < 0.001$.

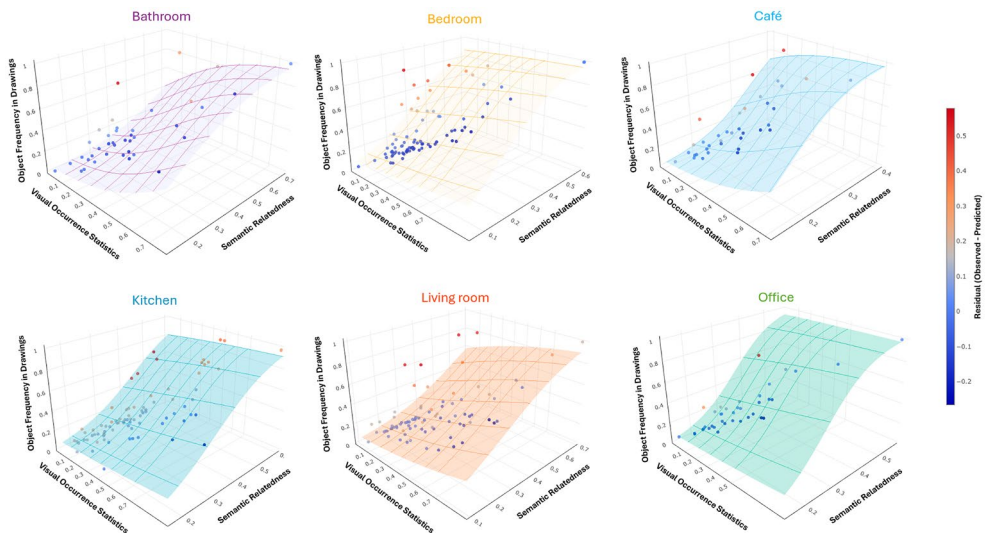


FIGURE 3 Visualization of generalized linear model fits when predicting object drawing frequencies from visual occurrence statistics and semantic relatedness, separately for the six categories. In all categories, both predictors yielded coefficients significantly greater than zero. Except for the bathroom and café category, the coefficient for visual occurrence statistics exceeds that for semantic relatedness, indicating a stronger influence of visual experience on drawing composition.

(i.e. how much more often does a chair appear in a living room compared to other scenes). Thus, we fitted mixed-effects GLMs that predicted object drawing frequency from visual and semantic specificity, again comparing single-factor to combined-factor models (Table 2). The results showed that the combined-factor model including both specificity predictors did not outperform the model that predicted drawing frequency from scene-specificity in visual occurrence statistics alone. This suggests that when selecting objects based on specificity, visual specificity (i.e. whether an object is more often seen in one room compared to other rooms) trumps semantic specificity (i.e. whether an object is semantically associated more strongly with one category than with the others).

However, it is worth noting that overall model performance in this analysis was relatively low, as reflected in the low R^2 values. This suggests that specificity alone might not drive object selection in drawings. Rather than selecting objects because they are highly specific to a scene, participants may instead choose objects based on how frequently they appear in a given contexts regardless of how often they appear elsewhere. This raises another question: for infrequently drawn objects, are people simply selecting objects that are broadly associated with many scenes and thus go well with

TABLE 2 Regression weights and goodness-of-fit statistics for generalized linear models with specificity predictors.

Scene category	Visual	Semantic	AIC	BIC	R ²
Full Model	0.55***	/	−343.73	−328.12	0.13
	/	1.16**	−328.38	−312.77	0.05
	0.56***	−0.05	−341.73	−322.22	0.13

** $p < 0.01$, *** $p < 0.001$.

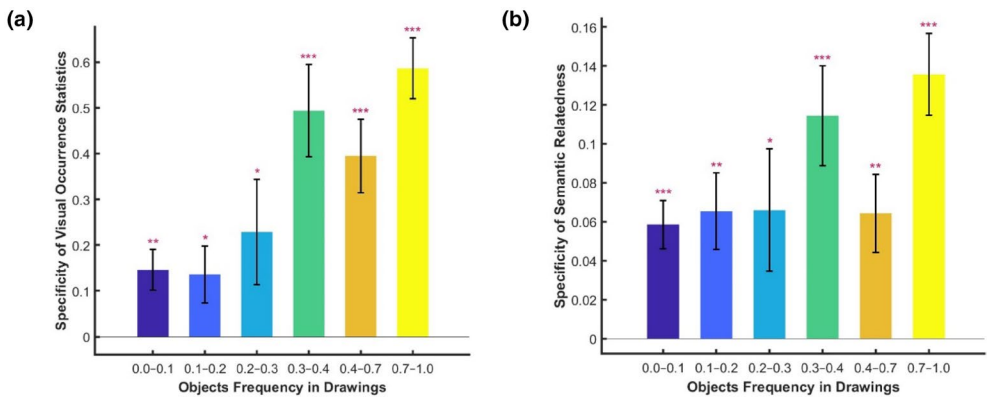


FIGURE 4 Assessing visual and semantic specificity of frequent and rare objects in scene drawings. Both (a) visual specificity and (b) semantic specificity are significantly positive across drawing frequency bins, suggesting that even rarely drawn objects are, on average, scene-specific. Error margins represent s.e.m. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

everything? Or were these infrequent objects still scene-specific? To address this, we utilized our specificity measure to examine whether even infrequently drawn objects were still specific for the scene categories they were drawn in. As expected, frequently drawn objects (those drawn in at least 40% of scenes) showed clear specificity both in visual occurrence statistics (all $t > 4$, all $p < 0.0001$) and semantic specificity (all $t > 3$, all $p < 0.01$). Critically, our analysis showed that even infrequent objects (those drawn in less than 40% of scenes) retained significant specificity across both visual (all $t > 2$, all $p < 0.05$) and semantic measures (all $t > 2$, all $p < 0.05$; Figure 4a,b), suggesting that infrequently included objects are still specifically associated with the target scene category (e.g. a hair dryer as a low-frequency object with high specificity for a bathroom), rather than simply being generic or universally compatible.

DISCUSSION

This study examined how visual occurrence statistics and semantic relatedness determine which objects are drawn when participants compose drawings of typical real-world scenes. Across six scene categories, we demonstrate that both factors significantly predicted how often objects were included in the drawings, and a combined model explained more variance than either predictor alone. Nonetheless, visual occurrence statistics consistently emerged as the stronger predictor. How often objects appear in scenes of a certain category and how strongly they are related to the scene concepts predicted object drawing frequencies better than the scene-specificity of the objects (i.e. whether an object is more often found in or more strongly related to the drawn scene category than to the other scene categories). Yet, even objects that were drawn infrequently were, on average, diagnostic for the scene category they were included in.

A natural question is what the semantic predictor adds beyond the visual one, given that semantic associations between scenes and objects are themselves probably grounded in visual experience

to a large extent. There are two reasons why semantic similarity offers additional predictive power. First, the visual predictor is necessarily incomplete. Our visual occurrence statistics were derived from ADE20K, which provides a large but bounded sample of real-world scenes. It cannot capture the full visual diet of any individual participant—including scenes encountered across different cultures and personal contexts. Therefore, variance that the semantic predictor explains may reflect visual statistics present in the real world but absent from ADE20K. Second, conceptual knowledge may incorporate non-visual sources. Object–scene associations stored in long-term memory are likely shaped not only by direct visual encounters but also by language, culture, and functional reasoning about what belongs in a scene. Semantics also allow humans to appreciate scenes they have never seen themselves (e.g. imagine a friend visiting a café during their holiday and describing it to you). Adjudicating between these accounts would require visual statistics derived from more comprehensive or individually tailored image samples, or controlled comparisons of objects that differ systematically in their visual versus semantic co-occurrence with scenes.

Producing a typical scene drawing necessitates the retrieval of relevant objects from long-term memory. Based on the classical schema theory (Biederman et al., 1982; Boyce et al., 1989), participants initially activate a semantic representation of the scene category from long-term memory. They then use prediction of expected semantic associations between the objects and scenes to guide visual information gathering (Bar, 2004; Leroy et al., 2020; Oliva & Torralba, 2007). These scene representations constrain the candidate objects that belong in the corresponding scene. The subsequent retrieval of objects is likely constrained by the structure of the scene, such as the typical spatial distributions of objects (Bar, 2004; Kaiser et al., 2019; Kaiser et al., 2015; Vö et al., 2019) and the spatial layout of whole scenes (Kaiser & Cichy, 2021). Critically, detailed representations of objects within scenes rely heavily on visual long-term memory (Brady et al., 2009), which is shaped by everyday visual statistical learning (Stansbury et al., 2013). Through repeated exposure to visual occurrence statistics in daily life, participants implicitly learn the typical composition and spatial arrangement of objects, enabling them to accurately predict and reconstruct detailed object information during the drawing process. Thus, both semantic and visual formats jointly contribute to the retrieval and representation of typical scenes. Interestingly, the observed dominance of visual occurrence statistics in predicting object frequency in drawings may reflect the fundamental role of real-world visual experience in shaping scene representations stored in long-term memory. Specifically, repeated visual encounters with objects in particular scene contexts likely strengthen their internal representations, making these frequently encountered objects more easily retrievable and visually detailed during reconstruction (Brady et al., 2009; Torralba et al., 2006).

On the other hand, the relatively weaker performance of our semantic predictor might partly stem from the hubness problem in word2vec-based semantic measurement (Schnabel et al., 2015). Specifically, when words are projected into high-dimensional vector spaces, ‘hubs’ appear as nearest neighbours to a disproportionately large number of other points (Radovanovic et al., 2010). For example, ‘kitchen’ becomes a hub that attracts objects to similar distances, which causes a narrow similarity range. This problem might compress the distribution of semantic associations, causing many scene-object pairs to tightly cluster within a restricted similarity band. Therefore, the low variance likely limits the measure's ability to explain frequency values. Future studies could apply models that yield more variability and thus diagnostic power, such as BERT-based (Devlin et al., 2019) and CLIP-based (Radford et al., 2021) measurements.

Furthermore, scene representations stored in memory are rooted in differences in participants' visual diets and may thus differ substantially between individuals (Engeser et al., 2025). In this study, we utilized object frequencies derived from the ADE20K image database and semantic associations from the word2vec model as proxies for real-world visual experience and conceptual knowledge. Despite essentially ignoring all inter-individual variance, this approach still predicted drawing content, highlighting the potential of this method to assess internal visual representations. Future research could incorporate more individualized measures, such as individual photographic exposure logs, personal digital archives or models trained on bespoke cultural and linguistic backgrounds. Such refinements might better characterize visual and semantic contributions on the individual or group level, thereby yielding more accurate predictions from visual and semantic factors.

Interestingly, the specificity of an object for a given scene category (i.e. whether an object was by itself diagnostic for the scene category) was a relatively weak predictor of object drawing frequency. This suggests that participants adopted a task-specific mindset in which they mainly focused on absolute object occurrence statistics that were best suited to maximize scene typicality (Wiesmann & Vö, 2023). If our task shifted from 'draw a typical bedroom' to 'draw the most diagnostic bedroom', object selection should pivot towards a greater weighting of specificity, increasing the contribution of items that are perhaps rarer but more exclusive to the category. It is also worth noting that our specificity measure was only computed relative to the five alternative categories in the current set. Future work could compute the scene-specificity of objects in relation to a broader set of reference categories to more accurately gauge specificity.

An additional contribution of this work is the public release of a large, object-annotated database of human typical scene drawings, the *Room Drawings Dataset*, which contains more than 1000 drawings across six categories (see Materials and Methods). Because each drawing comes with category and object labels and participant information, the dataset supports a wide range of secondary analyses, for instance, training and testing computational models on the drawings; comparing visual versus semantic predictors across populations; benchmarking typical scene construction in clinical or developmental cohorts; and linking typical drawing content to scene-selective neural responses. Furthermore, the dataset allows for easily appending new drawings, new scene categories, additional object codes or supplemental metadata. We hope this resource will serve as a useful resource for modelling scene perception and cognition.

The current study focused on object content and how it can be explained with visual occurrence statistics and semantic similarity measures. Yet, drawings also contain compositional information that is reflected in the absolute and relative positioning of the objects. Future studies could look at such compositional factors and readily use our open Room Drawings Dataset for this purpose. For example, the temporal order in which objects are drawn may itself be informative about how participants prioritise objects in a scene, which can in turn affect attentional prioritization in scene exploration (Engeser et al., 2026). Future work could exploit the temporal evolution of drawings, and combine these data with ratings of typicality or familiarity, to test which factors determine the order in which objects are drawn.

Another question concerns the extent to which typical scene drawings are shaped by participants' everyday environments. Future work could directly compare typical and personal scene representations within the same individuals—for example, by asking participants to draw both a typical version of a room and their own room, or to provide photographs of their own rooms alongside their typical drawings (Engeser & Kaiser, 2025). Such within-subject comparisons could inform how strongly personal experience influences typical scene composition and how this influence varies across scene categories and across participants with different living environments.

In sum, our findings reveal visual occurrence statistics and semantic relatedness jointly predict the composition of typical scene drawings. Visual frequency exerts the stronger influence on which core objects are rendered, and even infrequently drawn objects still contribute to category identity. These insights provide a new behavioural window onto the internal representations of the world that support human scene understanding across individuals.

AUTHOR CONTRIBUTIONS

Gongting Wang: Conceptualization; methodology; software; data curation; formal analysis; validation; investigation; writing – original draft; writing – review and editing. **Ilker Duymaz:** Methodology; software; investigation; writing – review and editing. **Matthew J. Foxwell:** Methodology; software; investigation; writing – review and editing. **Micha Engeser:** Methodology; software; investigation; writing – review and editing. **David Pitcher:** Methodology; software; investigation; writing – review and editing; funding acquisition. **Radoslaw M. Cichy:** Project administration; writing – review and editing; supervision; funding acquisition. **Daniel Kaiser:** Conceptualization; methodology; investigation; project administration; funding acquisition; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy (EXC 3066/1 'The Adaptive Mind', project no. 533717223). It was further supported by a European Research Council (ERC) Starting Grant (PEP, ERC-2022-STG 101076057). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors thank Daniela Marinova, Nasibeh Babaei and Pietra Pacheco Alves for their support in collecting the drawings. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at <https://osf.io/byu24/>.

ORCID

Gongting Wang  <https://orcid.org/0000-0002-2364-0639>

REFERENCES

- Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age Ageing*, 27, 399–403.
- Aubert, M., Brumm, A., Ramli, M., Sutikna, T., Saptomo, E. W., Hakim, B., & Dosseto, A. (2014). Pleistocene cave art from Sulawesi, Indonesia. *Nature*, 514(7521), 223–227.
- Bainbridge, W. A., & Baker, C. I. (2020). Boundaries extend and contract in scene memory depending on image properties. *Curr Biol*, 30, 537–543.e3.
- Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nat Commun*, 10(1), 5.
- Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci*, 5(8), 617–629.
- Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38(2), 347–358.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2), 143–177.
- Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 556–566.
- Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *J Exp Psychol Gen*, 138(4), 487–502.
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci*, 105(38), 14325–14329.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.
- Engeser, M., Ajith, S., Duymaz, I., Wang, G., Foxwell, M. J., Cichy, R. M., Pitcher, D., & Kaiser, D. (2025). Characterizing internal models of the visual environment. *Proc Biol Sci*, 292(2053), 20250602.
- Engeser, M., Babaei, N., & Kaiser, D. (2026). Can Individual Internal Models Predict Idiosyncratic Scene Exploration? *bioRxiv*, 2026–04.
- Engeser, M., & Kaiser, D. (2025). Inter-individual similarities in internal models of the world shape similarities in the perception and neural processing of scenes.
- Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nat Rev Psychol*, 2, 556–568.
- Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *J Appl Stat*, 31(7), 799–815.
- Fish, J., & Scrivener, S. (1990). Amplifying the mind's eye: Sketching and visual cognition. *Leonardo*, 23(1), 117–126.
- Goldschmidt, G. (2014). Modeling the role of sketching in design idea generation. In *An anthology of theories and models of design: Philosophy, approaches and empirical explorations* (pp. 433–450). Springer London.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
- Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences* (pp. 1833–1842). IEEE.

- Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., Alcolea-González, J. J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz, J., Weniger, G.-C., & Pike, A. W. G. (2018). U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science*, 359(6378), 912–915.
- Kaiser, D., & Cichy, R. M. (2021). Parts and wholes in scene processing. *J Cogn Neurosci*, 34(1), 4–15.
- Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends Cogn Sci*, 23(8), 672–685.
- Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychon Bull Rev*, 22(6), 1784–1790.
- Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, 34(1), 57–83.
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J Exp Psychol Gen*, 139(3), 558–578.
- Leroy, A., Faure, S., & Spotorno, S. (2020). Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. *Sci Rep*, 10(1), 8447.
- Long, B., Fan, J. E., Chai, Z., & Frank, M. C. (2019). Developmental changes in the ability to draw distinctive features of object categories. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 41). Cognitive Science Society.
- Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nat Commun*, 15(1), 1191.
- Metzger, W. (1936). *Gesetze des Sehens*. [Laws of vision.]. Kramer.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience*, 39(47), 9410–9423.
- Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends Cogn Sci*, 11(12), 520–527.
- Petherbridge, D. (2010). *The primacy of drawing*. Yale University Press.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PmLR.
- Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *J Mach Learn Res*, 11, 2487–2531.
- Roberts, B. R., & Wammes, J. D. (2021). Drawing and memory: Using visual production to alleviate concreteness effects. *Psychon Bull Rev*, 28(1), 259–267.
- Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015, September). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307). ACL.
- Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5), 1025–1034.
- 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.
- Vö, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210.
- Wang, G., Chen, L., Cichy, R. M., & Kaiser, D. (2025). Enhanced and idiosyncratic neural representations of personally typical scenes. *Proc Biol Sci*, 292(2043), 20250272.
- Wang, G., Foxwell, M. J., Cichy, R. M., Pitcher, D., & Kaiser, D. (2024). Individual differences in internal models explain idiosyncrasies in scene perception. *Cognition*, 245, 105723.
- Wechsler, D. (2009). *Wechsler memory scale* (4th ed.). Pearson.
- Wiesmann, S. L., & Vö, M. L. H. (2023). Disentangling diagnostic object properties for human scene categorization. *Sci Rep*, 13(1), 5912.
- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641). Computer Vision Foundation.
- Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2016). Semantic understanding of scenes through the ade20k dataset. arXiv preprint arXiv:1608.05442. <https://doi.org/10.48550/arXiv.1608.05442>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang, G., Duymaz, I., Foxwell, M. J., Engeser, M., Pitcher, D., Cichy, R. M., & Kaiser, D. (2026). How do visual and conceptual factors predict the object content in typical scene drawings? *British Journal of Psychology*, 00, 1–12. <https://doi.org/10.1111/bjop.70105>