# Complexity in Complexity: Understanding Visual Complexity Through Structure, Color, and Surprise

**Karahan Sarıtaş**[1]**, Peter Dayan**[1,2]**, Tingke Shen**[2*]**, Surabhi S Nath**[1,2,3*]
[1]University of Tübingen, Tübingen, Germany
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3]Max Planck School of Cognition, Leipzig, Germany

## Abstract

Understanding how humans perceive visual complexity is a key area of study in visual cognition. Previous approaches to modeling visual complexity assessments have often resulted in intricate, difficult-to-interpret algorithms that employ numerous features or sophisticated deep learning architectures. While these complex models achieve high performance on specific datasets, they often sacrifice interpretability, making it challenging to understand the factors driving human perception of complexity. Recently (Shen, et al. 2024) proposed an interpretable segmentation-based model that accurately predicted complexity across various datasets, supporting the idea that complexity can be explained simply. In this work, we investigate the failure of their model to capture structural, color and surprisal contributions to complexity. To this end, we propose Multi-Scale Sobel Gradient (MSG) which measures spatial intensity variations, Multi-Scale Unique Color (MUC) which quantifies colorfulness across multiple scales, and surprise scores generated using a Large Language Model. We test our features on existing benchmarks and a novel dataset (Surprising Visual Genome) containing surprising images from Visual Genome. Our experiments demonstrate that modeling complexity accurately is not as simple as previously thought, requiring additional perceptual and semantic factors to address dataset biases. Our model improves predictive performance while maintaining interpretability, offering deeper insights into how visual complexity is perceived and assessed. Our code, analysis and data are available at `https://github.com/Complexity-Project/Complexity-in-Complexity`.

## 1 Introduction

Visual complexity is a fundamental attribute of images that reflects the level of detail, intricacy, and variation in visual elements within a scene (Snodgrass & Vanderwart, 1980b). This perceptual characteristic encompasses multiple dimensions, including the number of lines within an icon or symbol, density of elements (McDougall et al., 1999; Forsythe et al., 2003), quantity of objects (Olivia et al., 2004; Saraee et al., 2020; Shin & Joo, 2019; Nath et al., 2024; Shen et al., 2024), clutter (Kyle-Davidson et al., 2023; Fan et al., 2017), symmetry (Kyle-Davidson et al., 2025), variety of colors (Corchs et al., 2016), and spatial organization of components (Yu & Winkler, 2013).

Understanding how humans perceive and judge visual complexity is crucial across numerous domains, from user interface design (Miniukovich et al., 2018; Akça & Ömer Özgür Tanriöver, 2021; Reinecke et al., 2013) to cognitive psychology (Forsythe, 2009; Madan et al., 2018). While deep learning models have achieved impressive performance in predicting perceived complexity, their black-box nature (Li et al., 2022) limits our ability to understand the underlying factors that drive human complexity judgments. This understanding is particularly important in applications where

---

[*]These authors contributed equally to the supervision of the project.

we need to not only predict but also explain or manipulate perceived complexity, such as in educational materials design (Stoesz et al., 2020; Ghai & Tandon, 2022) or information visualization (Zhu et al., 2007a;b). Moreover, interpretable features that correlate with human perception can provide valuable insights into cognitive processes, helping us better understand how the human visual system processes and evaluates complex scenes. Therefore, the development of simple, interpretable measures of visual complexity that align with human perception remains an important research goal, complementing the achievements of more complex computational approaches.

A recent study by Shen et al. (2024) bridged the gap between handcrafted features and deep learning approaches by proposing a simple two-feature model based on outputs from deep segmentation models (the numbers of segments and semantic classes in the image). Their simple, yet interpretable, model achieved superior performance compared to baseline on many naturalistic and art datasets, supporting the idea that the numbers of segments and classes are generic features that explain complexity well. However, their model fails to take into account the structure and arrangement of components in the image.

Here, we build upon the model from Shen et al. (2024), by studying two of its major failure modes. First, we find that datasets with high structural and color regularity require two low-level features to explain complexity: Multi-Scale Sobel Gradient (MSG) and Multi-Scale Unique Color (MUC). MSG captures continuous spatial intensity variations across multiple scales, offering a richer representation of image structure. MUC quantifies color diversity also at multiple scales, and at different color resolutions. The latter feature builds on the concept of "colorfulness" from (Teresa et al., 2014), and provides a more robust and performant measure of chromatic complexity.

We discovered a second failure mode of Shen et al. (2024) which concerns the rather underexplored contribution of whole-image, holistic or emergent information to visual complexity. Previous research by Forsythe et al. (2008) demonstrated that participants in a visual complexity experiment exhibited a familiarity bias, perceiving familiar shapes as less complex compared to unfamiliar or novel ones. Similarly, Snodgrass & Vanderwart (1980a) found a negative correlation between visual complexity and familiarity in their experiment. Forsythe (2009) suggests that complexity judgments are context-dependent and that familiarity should be considered when developing a model of visual complexity. Motivated by their findings, we show that holistic, whole-image surprise judgments of scenes contribute significantly to the visual complexity of naturalistic images. We define surprising images as those containing unusual, or contextually novel elements, i.e., the opposite of familiarity. (Meyer et al., 1997; Ortony & Partridge, 1987). To this end, we introduce a novel dataset, SVG, containing both surprising and randomly sampled images from the well-studied Visual Genome dataset (Krishna et al., 2016) and show that surprise scores generated by a large language model (LLM) explains significant variance in the perceived complexity of images in this dataset.

## 2 METHODS

### 2.1 MODELING AND EVALUATION

In this work, we evaluate the effectiveness of sets of features to explain visual complexity on datasets using linear regression models. We follow the procedure in Shen et al. (2024), and fit $M$ repetitions of 3-fold cross-validated linear regression. $M$ is determined by the size of the dataset to address statistical variability in smaller datasets. We measure performance using the mean Spearman correlation coefficient across all test splits. We compare the performance of our models to the same baselines in Shen et al. (2024): handcrafted features (Corchs et al., 2016; Kyle-Davidson et al., 2023) and deep learning approaches (Saraee et al., 2020; Feng et al., 2023). The code for our analysis, dataset [1] and experimental setup [2] can be found in our Github repositories.

### 2.2 DATASETS

We evaluate our method on four publicly available datasets with human-rated complexity scores: RSIVL (49 images) (Corchs et al., 2016), VISC (800 images) (Kyle-Davidson et al., 2023), SAVOIAS (1400 images across 7 categories) (Saraee et al., 2018), and IC9600 (9600 images across

---

[1]Code is available at: `https://github.com/Complexity-Project/Complexity-in-Complexity`
[2]Experimental setup is available at: `https://github.com/Complexity-Project/Experiment`

8 categories) (Feng et al., 2023). We select image subsets to demonstrate the effectiveness of our new features to explain complexity - VISC and IC9600 architecture subset for MSG, Savoias Art and Suprematism subsets for MUC, and Savoias Interior Design and IC9600 Abstract for the combination of MSG and MUC. Finally, we use our novel dataset SVG to show that surprisal contributes significantly to complexity.

## 2.3 SVG: A Dataset of Surprising Images

To the best of our knowledge, there is currently no existing image set with surprising images that can be used to study visual complexity. To fill this gap, we introduce a new dataset containing 100 highly surprising images and 100 (on average, less surprising) images, the latter of which were sampled uniformly at random from the Visual Genome, subject to stratification according to bins of complexity based on IC9600 scores (Feng et al., 2023). To confirm that our subjectively selected images are more surprising than randomly sampled ones, we compare the distributions of LLM-generated surprise scores (as explained in Section 2.4) between the two subsets. A histogram of these scores shows clear separation: sampled images cluster at lower values, while handpicked images have higher surprise scores (mean: 39.45 for ordinary, 72.35 for surprising). A Kolmogorov-Smirnov test confirms a significant difference ($D = 0.67, p < 0.001$), supporting our hypothesis. See Appendix A.3 for further details about the dataset, including representative examples.

We collect visual complexity ratings from humans using an online experiment on Prolific (Palan & Schitter, 2018). Participants are shown pairs of images and asked to select the one that is more visually complex, similarly to the procedure in Saraee et al. (2018). We followed the sampling strategy in Saraee et al. (2018), and collect $6000 = 200 \times 30$ comparisons by sampling images with probabilities inversely proportional to the total number of times they have already been selected. We ensure that each pairwise comparison is assessed by three different participants. Hence there are in total $18,000$ evaluations. Each participant evaluates 200 pairwise comparisons in one session. We also inserted attention checks into the experiment to ensure participant engagement and data reliability. We convert pairwise comparisons into scalar complexity ratings using the Bradley-Terry algorithm (Bradley & Terry, 1952). Initially, the final ratings were mostly clustered around zero, a phenomenon also observed in Saraee et al. (2018). To address this, we applied the same procedure as in their work. Before using the Bradley-Terry model, we rescaled the initial probability matrix from the range $[0, 1]$ to $[0.33, 0.66]$. This adjustment helped distribute the final ratings more evenly within a narrower and more interpretable range. We also collect human surprise ratings to validate our LLM-generated scores using a similar experiment. More details about our experiment setup can be found in Appendix A.2.

## 2.4 Generating Surprise Scores using LLMs

In order to have an automated pipeline for predicting complexity, we prompt an LLM using zero-shot Chain-of-Thought (Wei et al., 2023) to get surprise scores for each image in our SVG dataset. Our prompt is shown in Algorithm 1. We selected GEMINI-1.5-FLASH for its ability to provide rapid responses without compromising quality. The model generated surprise scores with step-by-step reasoning, producing structured output in the desired format. More details as to our preliminary ablation on LLMs can be found in Appendix A.4.

---

**Algorithm 1** Zero-shot-CoT

---

Q: Step by step, explain why this image is surprising or not. Consider factors like rare events, or unexpected content. Be precise in your reasoning. Then, on a precise scale from 0 to 100, rate the surprisal of this image.
Provide your reasoning and numeric rating as follows:
Reasoning: [your explanation]
Rating: ≪number≫

---

## 2.5 Multi-Scale Sobel Gradient

We propose the Multi-Scale Sobel Gradient (MSG) algorithm (Algorithm 2) to capture structure in images. MSG applies the Sobel operator across multiple resolutions to RGB images, functioning

as an asymmetry detector within $k \times k$ patches, where $k$ represents the kernel size. For symmetric patches, the left and right columns (or rows in horizontal application) of the kernel counterbalance each other, resulting in lower MSG values. Consequently, images exhibiting greater patch-level symmetry produce smaller MSG measurements. Note that the Sobel operator is typically applied to grayscale images (Kanopoulos et al., 1988), but we found in ablations that the grayscale version of the algorithm (which first converts the colored images to grayscale) did not perform as well. Based on our permutation tests across 16 datasets, the colored version demonstrated statistically significant superiority over the grayscale version in 6 datasets, while the grayscale version outperformed the colored in 2 datasets. The remaining 8 datasets showed no statistically significant differences between the two versions. Hence we use the color version of the algorithm for the rest of this work.

---

**Algorithm 2** Multi-Scale Sobel Gradient

---

**Require:** Input image $img \in \mathbb{R}^{H \times W \times 3}$ in RGB format
**Ensure:** Scalar MSG score
 1: Normalize image: $img \leftarrow img/255.0$
 2: Define scales $S = \{1, 2, 4, 8\}$
    and weights $W = \{0.4, 0.3, 0.2, 0.1\}$
 3: Initialize $MSG \leftarrow 0$
 4: **for** each scale $s \in S$ and weight $w \in W$ **do**
 5:    Resize image to $\frac{H}{s} \times \frac{W}{s}$:
       $scaled \leftarrow Resize(img, (\frac{H}{s}, \frac{W}{s}))$
 6:    **for** each channel $c \in \{R, G, B\}$ **do**
 7:       $G_x = Sobel(scaled[:, :, c], dx = 1, dy = 0)$
 8:       $G_y = Sobel(scaled[:, :, c], dx = 0, dy = 1)$
 9:       $mag = \sqrt{G_x^2 + G_y^2}$
10:       $grad_c \leftarrow Mean(mag)$
11:    **end for**
12:    $s\_grad \leftarrow Mean(grad_R, grad_G, grad_B)$
13:    $MSG \leftarrow MSG + w \cdot s\_grad$
14: **end for**
15: **return** $MSG$

---

### 2.6 MULTI-SCALE UNIQUE COLOR

We introduce MUC which intuitively counts the number of unique colors present in an image. We start from the "colorfulness" feature introduced by Teresa et al. (2014) (which was first applied to image indexing and content querying) and derive MUC by making colorfulness multi-scale. Note, Teresa et al. (2014) and subsequent works did not provide pseudocode for the colorfulness algorithm - we are the first to do so. MUC has two hyperparameters which control the coarse-graining of spatial resolution and color resolution (bit precision). In the rest of this paper we report results using the best bit precision for each dataset in terms of correlation to complexity (this was typically 7–8 bits). We use a fixed set of spatial scale weights for both MSG and MUC on all datasets.

## 3 RESULTS

The features we introduced: MSG, MUC, and Surprise, explain additional variance in perceived complexity beyond that captured by segmentation and class counts alone. We present these results in Table 1, showing that our models not only surpass all handcrafted features but also achieves performance comparable to the best supervised model. In the following subsections, we present each feature in turn and describe the datasets on which they improve complexity prediction. We then describe our final model, which combines our three novel features to achieve state-of-the-art or near state-of-the-art performance on all datasets.

### 3.1 EXPLAINING THE FAILURE MODES WITH A DATA-CENTRIC APPROACH:

---

**Algorithm 3** Multi-Scale Unique Color

---

**Require:** Image $I$ (RGB), Number of bits to preserve per channel $b$
**Ensure:** Multi-Scale unique color score
 1: Define scales $S = \{1, 2, 4, 8\}$
    and weights $W = \{0.4, 0.3, 0.2, 0.1\}$
 2: Initialize $MUC \leftarrow 0$
 3: **for** each scale $s \in S$ and weight $w \in W$ **do**
 4:    Resize image to $\frac{width}{s} \times \frac{height}{s}$:
       $I_s \leftarrow \text{resize}(I_{rgb}, (\frac{width}{s}, \frac{height}{s}))$
 5:    Calculate bit shift: $shift \leftarrow 8 - b$
 6:    Quantize image by bit shifting:
       $I_q \leftarrow (I_s \gg shift) \ll shift$
 7:    Flatten image: $I_{flat} \leftarrow \text{reshape}(I_q, -1, 3)$
 8:    Create color index:
       $idx \leftarrow I_{flat}[:, 0] \cdot 2^{16} + I_{flat}[:, 1] \cdot 2^8 + I_{flat}[:, 2]$
 9:    Count unique colors: $n_{\text{unique}} \leftarrow |\text{unique}(idx)|$
10:    Add weighted contribution:
       $MUC \leftarrow MUC + (w \cdot n_{\text{unique}})$
11: **end for**
12: **return** $MUC$

---

Table 1: Model performance on 7 datasets and comparison with previous models. The models from previous work are classified as being based on either handcrafted features or Convolutional Neural Networks (CNNs). * for supervised methods indicate their own *test* set. Bold indicates the best model. Improvements from the baseline are indicated with arrows.

| Model/Dataset | VISC | IC9. Arch. | Sav. Art | Sav. Suprematism | Sav. Int. | IC9. Abstract | SVG |
|---|---|---|---|---|---|---|---|
| **Handcrafted features** | | | | | | | |
| Corchs 1 (10 features) | 0.62 | 0.66 | 0.68 | 0.80 | 0.85 | 0.74 | 0.73 |
| Kyle-Davidson 1 (2 features) | 0.60 | 0.54 | 0.55 | 0.79 | 0.74 | 0.69 | 0.70 |
| **CNNs** | | | | | | | |
| Saraee (transfer) | 0.58 | 0.59 | 0.55 | 0.72 | 0.75 | 0.67 | 0.72 |
| Feng (supervised) | **0.72** | **0.92**\* | **0.81** | 0.84 | **0.89** | **0.94**\* | 0.83 |
| **Previous simple model** | | | | | | | |
| $\sqrt{num\_seg} + \sqrt{num\_class}$ | 0.56 | 0.66 | 0.73 | 0.89 | 0.61 | 0.66 | 0.78 |
| **Visual features** | | | | | | | |
| $MSG + MUC$ | 0.60 | 0.65 | 0.64 | 0.91 | 0.84 | 0.76 | 0.72 |
| **Baseline + visual features** | | | | | | | |
| $\sqrt{num\_seg} + \sqrt{num\_class} + MSG$ | 0.68 ↑0.13 | 0.76 ↑0.10 | 0.75 | 0.90 | 0.79 | 0.79 | 0.78 |
| $\sqrt{num\_seg} + \sqrt{num\_class} + MUC$ | 0.62 | 0.71 | **0.81** ↑0.08 | **0.94** ↑0.05 | 0.80 | 0.76 | 0.79 |
| $\sqrt{num\_seg} + \sqrt{num\_class} + MSG + MUC$ | 0.68 | 0.77 | 0.81 | 0.94 | 0.87 ↑0.26 | 0.83 ↑0.17 | 0.79 |
| **Baseline + semantic feature** | | | | | | | |
| $\sqrt{num\_seg} + \sqrt{num\_class} + Surprise$ | 0.60 | 0.67 | 0.74 | 0.89 | 0.60 | 0.67 | 0.83 ↑0.05 |
| **Baseline + all features** | | | | | | | |
| $\sqrt{num\_seg} + \sqrt{num\_class} + MSG + MUC + Surprise$ | 0.71 | 0.78 | **0.81** | **0.94** | 0.87 | 0.84 | **0.85** |

COLORS AND STRUCTURE

On the VISC dataset, the MSG feature improves the mean Spearman's rank correlation coefficient from 0.56 to 0.68, surpassing all previous handcrafted features. Similarly, it achieves a notable increase of 0.10 in mean Spearman's rank correlation coefficient on the IC9. Architecture dataset. These two datasets comprise of mostly real-world natural images of buildings, and indoor and outdoor scenes which are highly textured or regular - for example a grass field, the windows on a highrise, or rows of seats in an airport. A Sobel operator applied to a spatially coarse-grained image can be interpreted as an (a)symmetry detector. Hence MSG detects (a)symmetry and quantifies the regularity of local patches (e.g. windows on a highrise) at multiple spatial scales, making it a well-suited feature for predicting complexity on VISC and IC9. Architecture.

Table 1 shows that MUC improves model performance on Sav. Art dataset, increasing Spearman's correlation from 0.73 to 0.81, and in Sav. Suprematism, where it improves performance from 0.89 to 0.94. These datasets are comprised of 2D art. The efficacy of MUC in these domains likely stems from its ability to quantify color-based complexity, which may be more crucial for people's aesthetic judgments than judgments of naturalistic images.

(a) B: 50, P: 39, G: 40     (b) MSG: 20
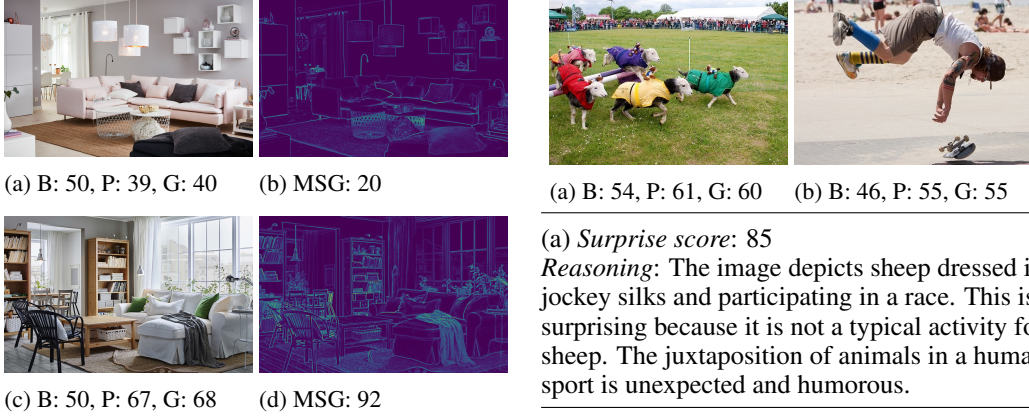
(c) B: 50, P: 67, G: 68     (d) MSG: 92

Figure 1: Left column: original images from *Sav. Int.*. Right column: gradient visualizations. B: baseline prediction using number of segmentations and classes. G = ground truth complexity. P = predicted complexity using baseline and MSG. All values are scaled between 0 and 100. The first image has 177 segmentations and 35 classes, while the second has 185 segmentations and 38 classes. Due to these similarities, the baseline model predicts nearly identical complexity scores for both. However, MSG acts as a latent dimension, refining predictions to better align with ground truth.



(a) B: 54, P: 61, G: 60     (b) B: 46, P: 55, G: 55

(a) *Surprise score*: 85
*Reasoning*: The image depicts sheep dressed in jockey silks and participating in a race. This is surprising because it is not a typical activity for sheep. The juxtaposition of animals in a human sport is unexpected and humorous.

(b) *Surprise score*: 85
*Reasoning*: The image depicts a skateboarder performing an unexpected maneuver on a beach, resulting in an unusual pose mid-air. The rarity stems from the seemingly uncontrolled backflip and the unusual beach setting for such a stunt.

Figure 2: Incorporating surprise scores (85) narrows the gap between baseline predictions and ground truth, highlighting surprise's role in aligning predictions with human perception. GEMINI-1.5-FLASH explanations improve the interpretability of these scores.

Two datasets, Sav. Interior Design and IC9. Abstract experience very large performance gains (from 0.61 to 0.87 and from 0.66 to 0.83 respectively) when both MSG and MUC features are included in the model. Interior Design is a dataset of synthetic renderings of rooms for IKEA advertisements. The dataset likely benefits from MSG for the same reason as VISC - the images contain regular, highly symmetric elements like books in a bookshelf or panels of a kitchen cabinet. Furthermore, unlike naturalistic images Sav. Interior Design images were designed with color themes in mind for aesthetic reasons, similar to art images which is likely why the Sav. Interior Design dataset benefits from MUC like the Sav. Art dataset. IC9. Abstract on the other hand, consists of real-life photographs. However, the images are taken in an artistic way with aesthetic value in mind, likely increasing the importance of colorfulness in predicting complexity.

Figure 1 presents two examples from the Sav. Interior Design dataset where the baseline model (Shen et al., 2024) assigns the same complexity due to similar SAM segmentations and FC-CLIP classes, despite differing ground truth complexities. However, the two images differ greatly in their MSG values (20 vs. 92 on a normalized scale of 100), possibly because the lower image has patterns on the bookshelves and blanket which contribute substantially to MSG. The lower (higher) MSG values in the top (bottom) image decrease (increase) the complexity predictions in model, making the predictions much closer to the ground truth complexity ratings.

## 3.2   SURPRISE TO THE RESCUE: CAPTURING THE COMPLEXITY BEYOND STRUCTURAL FEATURES

Table 1 shows that including surprise scores generated using GEMINI-1.5-FLASH improves Spearman's correlation on the SVG dataset from 0.78 to 0.83 over (Shen et al., 2024)'s baseline model. Our model containing surprise scores beats all the previous handcrafted features and matches the performance of (Feng et al., 2023)'s supervised network.

Figure 2 shows two examples from the SVG dataset where the model from (Shen et al., 2024) under-predicts visual complexity. Both images have high (85) surprise scores. Hence our model including

> **Participant Responses**
>
> **Question:** What strategy did you use to rate visual complexity?
>
> - "Checked if they look confusing or unreal."
> - "How common the situation was, the angle the image was taken at, the number of colours, the number of subjects and how in focus they were."
> - "How many different elements there were in each photo and how surprising or unusual the images were."
> - "The most complex were those with many different elements to them and the most surprising."
> - "The weird or uncanny images appeared quite complex to me, most of the landscape shots or regular street/traffic scenes didn't strike me as complex."
> - "Chose the confusing ones."
> - "Where there are lots of things to look at. Was there something at a distance in the background? Was the image novel?"
> - "The most complex were the ones with a lot going on in them—so more than one person or something unusual happening. The least complex were the ones of sport or vehicles which just look like generic images."
> - "If there was a lot going on in the image or if the image was out of the ordinary. Also, images where there were shadows and different types of light."
> - "I think it was a mix of structure of the image and the elements within it—the more unusual or juxtaposed the mix, the more complex it appeared to me."
> - "How much was going on in the image Also, whether things looked fake—e.g., the boy with the surfboard, where some looked real and some looked like computer-generated images. Then if things were incongruous (like things that weren't supposed to be there)."
> - "At first, I was thinking about the colours, people, and background. I then thought about how complex the act was in the photo."

Figure 4: A selection of responses from participants describing the strategies they used to assess visual complexity. Their responses highlight various factors such as the number of elements, unusual or weird features, level of detail, and how confusing or unreal an image appeared. Words related to surprise, such as 'unusual,' 'confusing,' and 'weird,' have been highlighted in red for emphasis.

surprise scores correctly increases its complexity predictions, making the predictions much closer to the ground truth complexity. Furthermore, one advantage of using LLMs in our automated pipeline for complexity prediction is LLMs offer (often) interpretable, explicit reasoning in natural language (Singh et al., 2024).
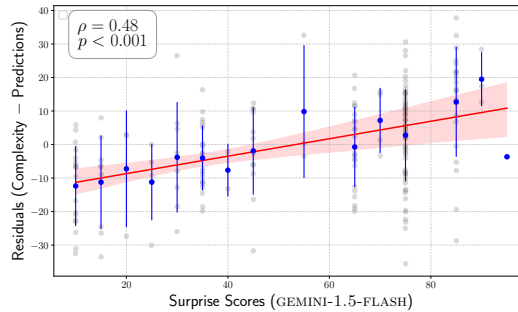


Figure 3: Correlation between residuals (actual complexity - baseline predictions) and surprise scores (SVG).

Two examples of explanations generated from GEMINI-1.5-FLASH using CoT prompting (Zhu et al., 2024; Wei Jie et al., 2024), can be seen in Figure 2. In the left image, GEMINI-1.5-FLASH successfully identifies the intuitively surprising fact that sheep are racing, an unusual activity for sheep. In the right image, GEMINI-1.5-FLASH identifies that it's unusual for the man to be doing an uncontrolled backflip, arguably consistent with human intuition for why the image is surprising.

Furthermore, Figure 3 shows that the surprise scores have a correlation of 0.48 ($p < 0.001$) with complexity after regressing out SAM number of segments and FC-CLIP number of objects, showing that surprise captures variance in complexity not explained by the number of segments or objects. These quantitative results are corroborated by participants' anecdotes on how they rated visual complexity on the SVG dataset. Many reported that "unusual", "surprising", or "weird", objects or events in the image, as well as judgments that the images are "confusing" or "unreal" caused them to rate the visual complexity higher. Figure 4 presents a selection of responses from participants in our experiment, in which they describe the specific strategies they used to assess visual complexity. Their responses provide a deeper understanding of the cognitive processes involved in evaluating image complexity and highlight how the *element of surprise* works as a latent dimension for visual complexity evaluation.

To explore quantitatively how well our GEMINI-1.5-FLASH surprise scores aligned with human judgments of surprise, we computed Spearman's correlation between GEMINI-1.5-FLASH scores and the human surprise scores we collected. The results revealed a strong positive correlation of 0.73, indicating substantial agreement between model-generated and human-derived assessments of visual surprise. This supports the fact GEMINI-1.5-FLASH explanations are good proxies for human judgments of image surprisal which contribute to human perceptions of visual complexity. In summary, GEMINI-1.5-FLASH surprise scores improve the predictive power of our complexity model on the SVG dataset and the natural language explanations generated alongside these scores facilitate a clearer understanding of why the image is complex.

The last row of Table 1 shows the performance of our final model including MSG, MUC, surprise as well as the two features from (Shen et al., 2024). Interestingly, we find combining MSG and MUC with surprise only improves performance on all datasets over using either of these features alone. Furthermore, our final model is significantly more performant than every baseline except (Feng et al., 2023)'s supervised network which is competitive. Results on all datasets using our final model can be found in Appendix A.1. Taken together, in contrast to what (Shen et al., 2024) previously suggested, the complexity of naturalistic images cannot be explained simply using two generic segment and object features. Some datasets require additional features to explain visual complexity fully.

Lastly, we evaluate the performance of GEMINI-1.5-FLASH on our SVG dataset using two distinct prompting strategies. In the first approach, the model estimated visual complexity *without* an explicit definition, replicating the open-ended methodology from our Prolific human-subject experiments. In the second approach, we provided the canonical definition of visual complexity as "the level of detail or intricacy contained within an image" (Snodgrass & Vanderwart, 1980a) to anchor the model's judgments. Both strategies leveraged Chain-of-Thought reasoning to generate structured rationales from GEMINI-1.5-FLASH. The resulting complexity ratings showed strong internal consistency (Spearman's $\rho = 0.89$ between the two strategies) and high alignment with human ground-truth ratings ($\rho = 0.83$ for both). Notably, our final integrated model—which combines pixel-level features, object-level segmentation metrics, and surprise scores from GEMINI-1.5-FLASH (quantifying how *surprising* an image is)—outperformed GEMINI-1.5-FLASH 's standalone visual complexity evaluations on the SVG dataset.

In Figure 5, we present another pair of images where, this time, the addition of surprise scores causes their predicted complexity values to diverge in opposite directions.

## 4 DISCUSSION

This work introduces three features to solve the inadequacy of the generic constructs proposed by (Shen et al., 2024) to explain complexity in certain datasets: multi-scale gradient analysis, multi-scale unique color, and a surprise score. We experimentally show that complexity can be multifaceted - requiring both generic features (segmentation and object counts) and dataset specific features to predict accurately. Our results demonstrate that pixel-level structural and chromatic features, MSG and MUC effectively capture key aspects of visual complexity, with MSG quantifying patch symmetry across scales and MUC capturing color diversity, both complementing object-level information from deep segmentation models. Additionally, we introduce surprise, an image-level "cognitive" feature, as a previously underexplored dimension of visual complexity, representing the degree to which the image as a whole deviates from expected objects or events. We introduce a new dataset SVG containing images with unexpected or novel elements to show that surprise contributes
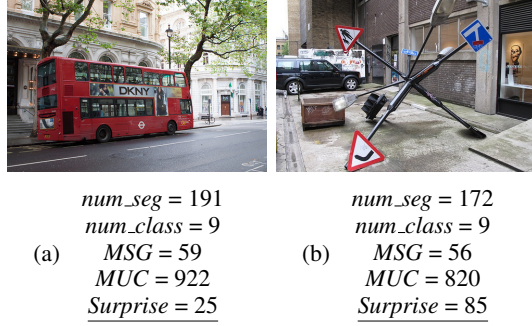
**Surprise as the hidden dimension**
We present a case where both images share similar visual features, including segmentations, classes, MSG, and MUC. The right image, with lower feature values, was initially predicted to have lower complexity. The left image received a complexity score of 60, while the right received 55. However, the element of surprise shifts the outcome, with improved predictions at 52 (72) compared to ground truth values of 25 (74). Despite their similarity, humans may perceive the second image as more complex possibly due to its unusual composition, where multiple traffic signs are clustered together.

Figure 5: Comparison of two images having similar values of visual features but with differing complexity evaluations.
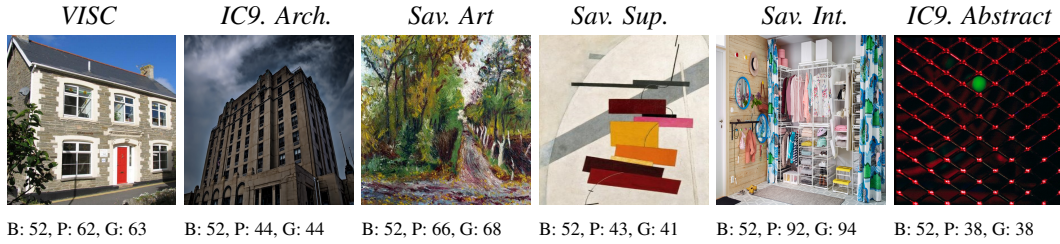


Figure 6: Representative samples from previous datasets show initial complexity predictions ($B = 52$), derived from $\sqrt{num\_seg}$ and $\sqrt{num\_class}$, refined using **MSG** (VISC, IC9 Architecture), **MUC** (Sav. Art, Sav. Suprematism), and **MSG + MUC** (Sav. Int. Design, IC9 Abstract). The enhanced predictions ($P$) align closer to ground truth scores ($G$).

significantly to perceived complexity. We constructed our SVG dataset by manually selecting surprising images and hence acknowledge the limitations in scalability and the potential subjectivity of this approach. We hope our preliminary yet confirmatory results will inspire future research to develop larger datasets using objectively defined measure to further validate the link between surprise and complexity.

**Numerosity Perception**    As the baseline approach, Shen et al. (2024) uses the square root of the number of segmentations and classes, an empirical finding that demonstrated better performance than using the raw values. One reason why the square root works better than the raw values can be explained by the Weber-Fechner Law (Moyer & Landauer, 1967), which establishes that subjective perception typically relates to stimulus intensity in a *nonlinear* manner. According to this principle, the just-noticeable difference between two stimulus intensities is proportional to the absolute stimulus intensity. From a practical perspective, this means participants can more readily distinguish between images containing 10 versus 20 segmentations than between those with 110 versus 120 segmentations, despite the absolute difference being identical. We also tested applying logarithmic transformation to the number of segmentations and classes, but the square-root transformation achieved slightly better results in the regression model, so we decided to stick with it.

**MSG versus Patch Symmetry and Edge Density**    Patch symmetry (Kyle-Davidson et al., 2023; Olivia et al., 2004; Rosenholtz et al., 2007) and edge density (Guo et al., 2018; Ciocca et al., 2015) are two commonly used low-level features for capturing structural information in an image. We run ablation studies to investigate whether MSG is more performant than these two features. We evaluate on all 16 datasets: VISC, RSIVL, SAVOIAS (5 categories - excluding Advertisement and Visualizations due to heavy presence of text), IC9600 (8 categories), and SVG using permutation tests. We compare MSG against patch symmetry because (Shen et al., 2024) reported that it captures structural information well on the VISC dataset. We compare MSG against Canny edge density because Canny is a standard metric for predicting visual complexity. It has been applied in Guo et al. (2018); Corchs et al. (2016); Rosenholtz et al. (2007). You can find more details of our ablation study in Appendix A.5.

MSG performs significantly better than Canny edge density in five dataset (IC9. Abstract, IC9. Paintings, Sav. Scenes, IC9. Advertisement, and Sav. Interior Design). For example, on IC9. Abstract, MSG achieves significantly higher correlation with the complexity compared to edge density (0.517 vs 0.705). Edge density only outperformed MSG on two datasets (IC9. Person and Sav. Art). Compared to patch symmetry, MSG performs better on 7 datasets (IC9. Abstract, IC9. Paintings, IC9. Scenes, IC9. Advertisement, IC9. Architecture, IC9. Person, and Sav. Art). Patch symmetry outperforms MSG on only one dataset (Sav. Interior Design dataset). Our ablations show that MSG is currently the most performant structural feature for predicting complexity, outperforming both edge density and patch symmetry across diverse image datasets.

**Single versus Multi-Scale** We perform ablation studies to investigate the necessity of multi-scale computations for MSG and MUC. We measure the correlation between complexity and our visual features using both single and multiple scales. By single scale, we refer to the application of our algorithms on the whole image only once ($W = 1.0$ and $S = 1$) without rescaling. We find that multi-scale MSG outperforms single-scale MSG on 8 datasets (VISC, IC9 Abstract, Sav. Objects, Sav. Scenes, Sav. Suprematism, IC9 Transport, IC9 Objects, and SVG), while single-scale MSG outperforms multi-scale MSG on just two datasets (IC9 Paintings and IC9 Architecture) out of the 16 datasets tested. The advantage of the multi-scale approach was particularly pronounced in Sav. Suprematism, where correlation increased substantially from 0.474 to 0.616. Similarly, multi-scale MUC outperformed single-scale MUC on 5 datasets (IC9. Abstract, IC9. Paintings, Sav. Art, IC9. Advertisement, IC9. Architecture), while single-scale MUC outperformed multi-scale MUC on only one dataset (IC9 Transport). The multi-scale versions of MSG and MUC maximize the correlation with the perceived visual complexity.

**Memorability and Aesthetics** The existing literature on visual complexity has largely concentrated on identifying low-level image features that define complexity, rather than examining its perceptual implications for the observer. However, as noted by (Forsythe, 2009), complexity inherently relates to the difficulty of cognitive processing, suggesting that surprising images should be perceived as more complex. We provide evidence in support of this hypothesis, showing that surprise, as determined by an LLM, contributes significantly to visual complexity on our SVG dataset.
Prior work links image memorability, a closely related concept to complexity (Kyle-Davidson & Evans, 2023), to distinctiveness, finding that unique or atypical features enhance retention. (Bruce et al., 1994; Bartlett et al., 1984) find that distinct facial features are more memorable. (Lukavský & Děchtěrenko, 2017) find that images that deviate from their neighbors in a CNN embedding space have increased recall. We observe a weak positive correlation between human surprisal ratings and both AMNet-predicted ($\rho = 0.25, p < 0.05$) (Fajtl et al., 2018) and ResMem-predicted memorability ($\rho = 0.25, p < 0.05$) (Needell & Bainbridge, 2022), suggesting that surprise may contribute to, or even be a common cause for both judgments of complexity and memorability. Future works should elucidate the relationship between complexity, memorability, and surprise.
The relationship between complexity and aesthetics has been extensively studied, though its precise functional form remains uncertain. (Imamoglu, 2000; Berlyne, 1970) provides evidence for an inverted U-shape between visual complexity and aesthetics. (Sun et al., 2015) confirms the ascending part of Berlyne's inverted-U curve and employs complexity features to predict beauty judgments in images. However, their measure relies solely on visual features like edges and symmetry, while the authors acknowledge that semantic factors, such as familiarity, may also shape perceived visual complexity. In this work, we find preliminary evidence that surprise contributes to complexity. Given the close relationship between complexity and aesthetics, it would be interesting in future works to investigate the extent to which surprise also influences aesthetic, potentially by mediating complexity. Our findings emphasize the multifaceted nature of complexity, requiring varied features across datasets. An incomplete representation of visual complexity may obscure its link to aesthetics, while a more comprehensive approach integrating visual and cognitive-semantic features could enhance understanding.

**Future work** While our work uses whole-image, holistic surprise scores, GEMINI-1.5-FLASH identifies specific object relationships and object-contexts that make each image more or less surprising. Therefore, it would be interesting for future works to derive surprise scores (or even other semantic features) from object-centric representations such as scene graphs. Our dataset, SVG enables this type of investigation by including human-rated complexity scores, and inheriting region descriptions, and scene graphs from the Visual Genome dataset. These rich annotations open a promising path for developing and validating features that reflect semantic rather than purely visual contributions to complexity.

REFERENCES

Eren Akça and Ömer Özgür Tanriöver. A comprehensive appraisal of perceptual visual complexity analysis methods in gui design. *Displays*, 69:102031, 2021. ISSN 0141-9382. doi: https://doi.org/10.1016/j.displa.2021.102031.

J. C. Bartlett, S. Hurry, and W. Thorley. Typicality and familiarity of faces. *Memory & Cognition*, 12(3):219–228, 1984. doi: 10.3758/BF03197669.

D. E. Berlyne. Novelty, complexity, and hedonic value. *Perception & Psychophysics*, 8:279–286, 1970. doi: 10.3758/BF03212593.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029. URL https://doi.org/10.2307/2334029. Accessed 21 Jan. 2025.

V. Bruce, A. M. Burton, and N. Dench. What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 47(1):119–141, 1994. doi: 10.1080/14640749408401146.

Gianluigi Ciocca, Silvia Corchs, and Francesca Gasparini. Complexity perception of texture images. In Vittorio Murino, Enrico Puppo, Diego Sona, Marco Cristani, and Carlo Sansone (eds.), *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, pp. 119–126, Cham, 2015. Springer International Publishing. ISBN 978-3-319-23222-5.

Silvia E Corchs, Gianluigi Ciocca, Emanuela Bricolo, and Francesca Gasparini. Predicting complexity perception of real world images. *PLOS ONE*, 11(6):e0157986, 2016. doi: 10.1371/journal.pone.0157986.

J.R. de Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47:1–12, 2015. doi: 10.3758/s13428-014-0458-y. URL https://doi.org/10.3758/s13428-014-0458-y.

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention, 2018. URL https://arxiv.org/abs/1804.03115.

Zhen Bao Fan, Yi-Na Li, Jinhui Yu, and Kang Zhang. Visual complexity of chinese ink paintings. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351485. doi: 10.1145/3119881.3119883. URL https://doi.org/10.1145/3119881.3119883.

Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–17, 2023. doi: 10.1109/TPAMI.2022.3232328.

Alex Forsythe, Noel Sheehy, and Martin Sawey. Measuring icon complexity: an automated analysis. *Behavior Research Methods, Instruments, & Computers*, 35(2):334–342, 2003. doi: 10.3758/bf03202562.

Alex Forsythe, Gerry Mulhern, and Martin Sawey. Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior research methods*, 40:116–29, 03 2008. doi: 10.3758/BRM.40.1.116.

Alexandra Forsythe. Visual complexity: Is that all there is? In Don Harris (ed.), *Engineering Psychology and Cognitive Ergonomics*, pp. 158–166, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-02728-4.

Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Akanksha Ghai and U. Tandon. Analyzing the impact of aesthetic visual design on usability of e-learning: An emerging economy perspective. *Higher Learning Research Communications*, 12: 1–22, 01 2022. doi: 10.18870/hlrc.v12i2.1325.

Xiaoying Guo, Yuhua Qian, Liang Li, and Akira Asano. Assessment model for perceived visual complexity of painting images. *Knowledge-Based Systems*, 159:110–119, 2018. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.06.006.

Imamoglu. Complexity, liking and familiarity: Architecture and non-architecture turkish students' assessments of traditional and modern house facades. *Journal of Environmental Psychology*, 20 (1):5–16, 2000. ISSN 0272-4944. doi: https://doi.org/10.1006/jevp.1999.0155.

N. Kanopoulos, N. Vasanthavada, and R.L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988. doi: 10.1109/4.996.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

Cameron Kyle-Davidson and Karla K Evans. Complexity & memorability have a nonlinear relationship when remembering scenes. *Journal of Vision*, 23(9):5251–5251, 2023.

Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B. Walther, Adrian G. Bors, and Karla K. Evans. Characterising and dissecting human perception of scene complexity. *Cognition*, 231:105319, 2023. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2022.105319.

Cameron Kyle-Davidson, Oscar Solis, Stephen Robinson, Ryan Tze Wang Tan, and Karla K. Evans. Scene complexity and the detail trace of human long-term visual memory. *Vision Research*, 227: 108525, 2025. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2024.108525.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond, 2022. URL https://arxiv.org/abs/2103.10689.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiří Lukavský and Filip Děchtěrenko. Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception & Psychophysics*, 79(7):2044–2054, 2017. doi: 10.3758/s13414-017-1375-9.

Christopher R Madan, Janine Bayer, Matthias Gamer, Tina B Lonsdorf, and Tobias Sommer. Visual complexity and affect: Ratings reflect more than meets the eye. *Frontiers in Psychology*, 8:2368, 2018. doi: 10.3389/fpsyg.2017.02368.

Sine J.P. McDougall, Martin B. Curry, and Oscar de Bruijn. Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers*, 31(3):487–519, 1999. doi: 10.3758/BF03200730.

Wulf U. Meyer, Rainer Reisenzein, and Achim Schützwohl. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21(3):251–274, 1997. doi: 10.1023/A:1024422330338. URL https://doi.org/10.1023/A:1024422330338.

Aliaksei Miniukovich, Simone Sulpizio, and Antonella De Angeli. Visual complexity of graphical user interfaces. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356169. doi: 10.1145/3206505.3206549. URL https://doi.org/10.1145/3206505.3206549.

Robert S. Moyer and Thomas K. Landauer. Time required for judgements of numerical inequality. *Nature*, 215:1519–1520, 1967. doi: 10.1038/2151519a0.

Surabhi S Nath, Franziska Brändle, Eric Schulz, Peter Dayan, and Aenne Brielmann. Relating objective complexity, subjective complexity, and beauty in binary pixel patterns. *Psychology of Aesthetics, Creativity, and the Arts*, 2024.

Coen D. Needell and Wilma A. Bainbridge. Embracing new techniques in deep learning for estimating image memorability, 2022. URL https://arxiv.org/abs/2105.10598.

Aude Olivia, Michael L. Mack, Mochan Shrestha, and Angela S. Peeper. Identifying the perceptual dimensions of visual complexity of scenes. 2004. URL https://api.semanticscholar.org/CorpusID:8049814.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Andrew Ortony and Derek Partridge. Surprisingness and expectation failure: What's the difference? pp. 106–108, 01 1987.

Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2017.12.004.

Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z. Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 2049–2058, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481281. URL https://doi.org/10.1145/2470654.2481281.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of vision*, 7 2: 17.1–22, 2007. URL https://api.semanticscholar.org/CorpusID:10847280.

Elham Saraee, Mona Jalal, and Margrit Betke. Savoias: A diverse, multi-category visual complexity dataset, 2018. URL https://arxiv.org/abs/1810.01771.

Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195:102949, 2020. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2020.102949.

Tingke Shen, Surabhi S Nath, Aenne Brielmann, and Peter Dayan. Simplicity in complexity : Explaining visual complexity using deep segmentation models, 2024. URL https://arxiv.org/abs/2403.03134.

Yooncheol Shin and Jaewoo Joo. Less for more, but how & why? – number of elements as key determinant of visual complexity. 09 2019.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models, 2024. URL https://arxiv.org/abs/2402.01761.

J. G. Snodgrass and M. Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):174–215, 1980a. doi: 10.1037//0278-7393.6.2.174.

Joan Gay Snodgrass and Mary Louise Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, 6 2:174–215, 1980b. URL https://api.semanticscholar.org/CorpusID:16871046.

Brenda Stoesz, Mehdi Niknam, and Jessica Sutton. Defining the visual complexity of learning management systems using image metrics and subjective ratings. *Canadian Journal of Learning and Technology / La revue canadienne de l'apprentissage et de la technologie*, 46, 12 2020. doi: 10.21432/cjlt27899.

Litian Sun, Toshihiko Yamasaki, and Kiyoharu Aizawa. Relationship between visual complexity and aesthetics: Application to beauty prediction of photos. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother (eds.), *Computer Vision - ECCV 2014 Workshops*, pp. 20–34, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16178-5.

Artese Teresa, Gianluigi Ciocca, and Isabella Gagliardi. Good 50×70 project: A portal for cultural and social campaigns. *Archiving Conference*, 11, 06 2014. doi: 10.2352/issn.2168-3204.2014. 11.1.art00048.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2148–2164. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.138.

Honghai Yu and Stefan Winkler. Image complexity and spatial information. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 12–17, 2013. doi: 10.1109/QoMEX.2013.6603194.

Ying Zhu, Xiaoyuan Suo, and G. Owen. Complexity analysis for information visualization design and evaluation. volume 4841, pp. 576–585, 11 2007a. ISBN 978-3-540-76857-9. doi: 10.1007/978-3-540-76858-6\_56.

Ying Zhu, Xiaoyuan Suo, and G. Scott Owen. Complexity analysis for information visualization design and evaluation. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Nikos Paragios, Syeda-Mahmood Tanveer, Tao Ju, Zicheng Liu, Sabine Coquillart, Carolina Cruz-Neira, Torsten Müller, and Tom Malzbender (eds.), *Advances in Visual Computing*, pp. 576–585, Berlin, Heidelberg, 2007b. Springer Berlin Heidelberg. ISBN 978-3-540-76858-6.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. Explanation in the era of large language models. In Rui Zhang, Nathan Schneider, and Snigdha Chaturvedi (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pp. 19–25, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-tutorials.3. URL https://aclanthology.org/2024.naacl-tutorials.3/.

# A APPENDIX

## A.1 RESULTS ON ALL DATASETS

Table 2 compares the performance of the baseline approach, our final model, and Feng's supervised model across all datasets.

## A.2 ADDITIONAL EXPERIMENT DETAILS

In our main paper, we briefly described the creation of our Surprising Visual Genome (SVG) dataset and our methodology for collecting human complexity ratings. Here, we provide additional details about our experimental procedure.

The visual complexity assessment task was conducted using Prolific (Palan & Schitter, 2018) as the participant recruitment platform. Each participant completed 200 pairwise comparisons in a single session. The study, which took approximately 15 minutes to complete, offered payment of 2.61 pounds. All participants were recruited from the UK.

The experiment was implemented using jsPsych (de Leeuw, 2015), a JavaScript library for creating behavioral experiments in web browsers. After reading instructions and completing practice trials, participants were presented with image pairs and asked to select the one they perceived as more visually complex (see Figure 7). Initial instructions are provided in Frame A.2.

Table 2: Correlations with the ground truth perceived complexity ratings. Bold indicates the best model performance for each dataset. Improvements from the baseline are indicated with arrows.

| Dataset | Baseline | *Our final model* | Feng (sup.) |
|---|---|---|---|
| VISC | 0.56 | 0.71 ↑ 0.15 | **0.72** |
| RSIVL | 0.83 | **0.84** ↑ 0.01 | 0.83 |
| SVG | 0.78 | **0.85** ↑ 0.07 | 0.83 |
| Sav. Art | 0.73 | **0.82** ↑ 0.09 | **0.82** |
| Sav. Sup. | 0.89 | **0.94** ↑ 0.05 | 0.84 |
| Sav. Int. | 0.61 | 0.87 ↑ 0.26 | **0.89** |
| Sav. Objects | 0.82 | **0.85** ↑ 0.03 | 0.83 |
| Sav. Scenes | 0.77 | 0.81 ↑ 0.04 | **0.87** |
| IC9. Arch. | 0.66 | 0.77 ↑ 0.11 | **0.93** |
| IC9. Abstract | 0.66 | 0.84 ↑ 0.18 | **0.95** |
| IC9. Scenes | 0.77 | 0.83 ↑ 0.06 | **0.93** |
| IC9. Objects | 0.80 | 0.84 ↑ 0.04 | **0.92** |
| IC9. Transport | 0.74 | 0.79 ↑ 0.05 | **0.88** |
| IC9. Paintings | 0.83 | 0.87 ↑ 0.04 | **0.95** |
| IC9. Adv. | 0.75 | 0.82 ↑ 0.07 | **0.93** |
| IC9. Person | 0.59 | 0.66 ↑ 0.07 | **0.85** |

Here we compare the baseline approach ($\sqrt{num\_seg} + \sqrt{num\_class}$), our final model (baseline $+ MSG + MUC + Surprise$), and the best supervised model (Feng et al., 2023) across 16 datasets. When evaluating on datasets not included in Feng's training data, our model outperforms Feng's supervised approach on 4 out of 8 datasets, while Feng's model performs better on 3 datasets (with one dataset resulting in a tie), giving us a final record of 4/3/1. Feng's supervised method shows superior performance across all IC9 datasets as expected since their model was specifically trained on this data distribution (in-domain evaluation). In contrast, our approach maintains strong performance while using interpretable features that don't require training on any dataset.

---

Thank you for your consent to participate in our study.

In this study, you will be shown pairs of images and asked to judge which one appears more **visually complex** to you.

You will make around 200 comparisons in total. For each pair, simply click on the image you judge to be more surprising. There are no right or wrong answers - we are interested in your perception of surprise.

At the end of the study, we will ask brief questions to better understand your judgements of visual complexity.

The complete study will take approximately 15 minutes.

**We will now show you some example pairs to familiarize you with the task. Click "Start" to begin.**

---

To ensure data quality, we embedded three attention check trials among the 200 comparisons. In these trials, participants were shown identical images, except one image contained a text overlay with explicit instructions to select it. Participants who failed more than one attention check were excluded from the analysis and replaced with new participants to maintain our target sample size.

After completing all comparisons, participants responded to questions about their strategy for rating visual complexity, whether they found particular types of images consistently more complex, and if they had additional comments about their experience. These responses provided qualitative insights into the factors influencing complexity judgments, as highlighted in the main paper.

The pairwise comparison data was processed using the Bradley-Terry algorithm (Bradley & Terry, 1952) to convert the binary choices into continuous complexity scores. To address the issue of scores clustering around zero, we applied the rescaling procedure described in Saraee et al. (2018), which involved transforming the probability matrix from [0,1] to [0.33, 0.66] before applying the Bradley-Terry model.

To collect human surprise ratings, we utilized the same experimental setting with minor modifications to the instructions and text to focus on surprisal rather than complexity.

Our complete experimental materials, including stimuli, code for the web-based interface, data processing scripts, and analysis pipelines, are available at `https://github.com/Complexity-Project/Experiment` to facilitate replication and extension of our work.

**Please look carefully at the following two images and decide which one is more visually complex.**
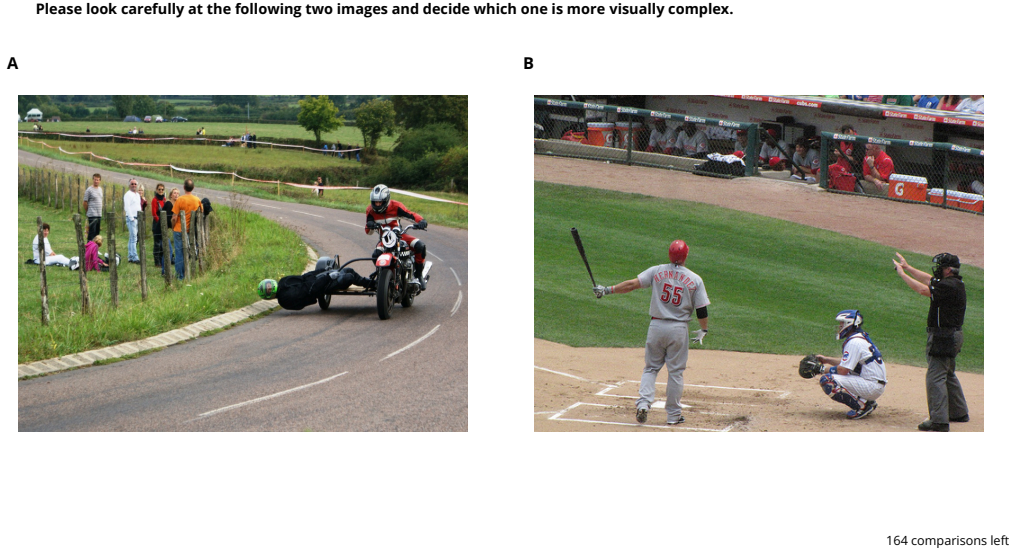


164 comparisons left

Figure 7: Interface of the pairwise comparison experiment for visual complexity assessment. Participants were instructed to select the image they perceived as more visually complex by clicking on it. The number of comparisons remaining was displayed in the bottom right corner.

## A.3 SVG DATASET

To curate the SVG dataset, we employed a targeted approach, first manually identifying candidates with high surprisal value from the Visual Genome dataset, then validating our selections using an automated method (using an LLM). We considered an alternative strategy—directly evaluating surprise values across the entire corpus using an LLM and selecting those exceeding a threshold—but determined this was impractical for the following: Identifying truly distinctive surprising images, as exemplified in Figure 8 and subsequently confirmed by participant responses in Figure 4, required examining a substantial portion of the dataset, which consists of $\geq$ 100K images. The computational demands of LLM processing at this scale imposed substantial practical constraints: either (1) lengthy processing times due to API rate limits, or (2) prohibitive costs for premium API access that would bypass these limitations. Therefore, we implemented a hybrid approach, first manually selecting candidate images with high surprisal potential, followed by LLM verification to ensure selection validity.



G: 15, H: 4, C: 5    G: 35, H: 32, C: 81    G: 65, H: 66, C: 27    G: 75, H: 76, C: 56    G: 85, H: 98, C: 72    G: 85, H: 85, C: 81

Figure 8: Representative samples from the SVG dataset. G indicates the GEMINI-1.5-FLASH-generated surprise scores, H represents human surprise scores collected through our experiments, and C shows the ground truth perceived visual complexity - all values are normalized to a 0-100 scale. The second figure illustrates cases where relatively ordinary images tend to be classified as visually complex merely due to their high number of objects or segmentations. Figures 4 and 5 share similar visual features (number of segmentations = 88 vs. 88, number of classes = 5 vs. 4, MSG = 37 vs. 38, MUC = 34 vs. 24), yet their visual complexity values differ significantly (56 vs. 72). This difference may be attributed to how participants perceived the latter image as more surprising (76 vs. 98).

16

## A.4 Ablations on LLMs for Surprise Scores

To generate surprise scores for images in our SVG dataset, we evaluated three different LLMs: GEMINI-1.5-FLASH (Gemini Team, 2024), LLAMA-3.2-8B (Llama Team, 2024), and CHATGPT-4O (OpenAI, 2024). Our selection strategy aimed to compare models across the accessibility spectrum: LLAMA-3.2-8B as a fully open-source model with unrestricted access, CHATGPT-4O as a sophisticated commercial model available only through paid API (or free account using the UI), and GEMINI-1.5-FLASH as a middle-ground solution offering free access with daily request limitations.

We input the exact same Chain-of-Thought prompt (as given in Algorithm 1) to the models. Although a popular choice as being fully open-source, LLAMA-3.2-8B could not keep up with the other models, as it struggled with understanding the task and providing responses that could be parsed by our algorithm. Therefore, we decided to focus on comparing GEMINI-1.5-FLASH and CHATGPT-4O for our task. In Table 3, we present the Spearman correlation coefficients between our predicted values and ground truth perceived visual complexity, comparing surprise scores derived from GEMINI-1.5-FLASH, CHATGPT-4O, and human participants as features in our regression models.

Table 3: Correlations of our models with the ground truth perceived visual complexity.

| Other features included | Surprise scores | | |
| --- | --- | --- | --- |
| | CHATGPT-4O | GEMINI-1.5-FLASH | Human-rated |
| $\sqrt{num\_seg} + \sqrt{num\_class}$ | 0.81 | 0.83 | **0.85** |
| $\sqrt{num\_seg} + \sqrt{num\_class} + \text{MSG} + \text{MUC}$ | 0.82 | 0.85 | **0.86** |

As anticipated, human-rated surprise scores yield the strongest contribution to the baseline model, enhancing correlation from 0.78 to 0.86. GEMINI-1.5-FLASH and CHATGPT-4O-generated surprisal scores show substantial agreement with a correlation of 0.78. However, the GEMINI-1.5-FLASH-generated surprise scores demonstrate superior performance compared to those produced by CHATGPT-4O, with the latter failing to surpass Feng's supervised model (0.83). Based on these findings, we ultimately selected GEMINI-1.5-FLASH-produced surprise scores for our final model—a choice that offered the additional practical advantage of free access within certain usage limits.

## A.5 Further Details on MSG and MUC

Here we present two examples in Figure 9 from the previous study (Shen et al., 2024), presented as failure modes.

We performed multiple ablation studies across all 16 datasets to determine the superior methodology. Our statistical analysis assessed whether the correlation between feature $X$ and visual complexity differs significantly from that between feature $Y$ and complexity, allowing us to identify the more effective approach.

The permutation test first calculates Spearman's rank correlations ($r_s$) between complexity $C$ and both features $X$ and $Y$, computing the difference of their absolute values, denoted as $\Delta_{\text{obs}} = |r_s(C,X)| - |r_s(C,Y)|$. It then performs $n = 1000$ permutations, each time randomly shuffling the complexity values $C^{(i)}$ while keeping feature values fixed, which breaks any real relationships between complexity and the features. For each permutation $i$, it calculates new correlations with the shuffled complexity values and records their permuted difference as $\Delta_{\text{perm}}^{(i)} = |r_s(C^{(i)},X)| - |r_s(C^{(i)},Y)|$. Statistical significance is determined by calculating a p-value:

$$p = \frac{\sum_{i=1}^{n} \mathbb{I}(|\Delta_{\text{perm}}^{(i)}| \geq |\Delta_{\text{obs}}|) + 1}{n+1}$$

MSG applies the Sobel operator to each $k \times k$ patch in the image. While the Sobel operator has symmetric magnitude, its kernel contains opposite signs on different sides. Consequently, symmetrical image patches typically produce lower MSG values. This image has a high segmentation count (538 - maximum in VISC dataset), yielding a high baseline prediction of 86 versus the ground truth of 56. Incorporating its MSG score of 0.26 (50th percentile) reduces the predicted complexity to 0.69, aligning closer to the actual value.



Similarly, this image receives a high baseline prediction (57) due to its substantial segmentation count (204, 78.5 percentile) and class diversity (37, 87 percentile). However, by incorporating both symmetry effects and color uniformity, we can refine our prediction. The image's low MSG value (0.19, 15th percentile) and MUC value (0.18 for 7 bits, 7th percentile) adjust the final prediction to 0.20, much closer to the ground truth complexity of 0.16. Both the symmetrical patches and monochromatic structure reduce its perceived visual complexity.

Figure 9: Examples where MSG and MUC features help improve complexity predictions: (top) an image from VISC dataset and (bottom) an image from Sav. Int. dataset, both previously shared by Shen et al. (2024) as failure modes.

| Feature | Comparison | Result (win/loss/non-significant) |
|---|---|---|
| MSG | *baseline* vs. *alternative #1* | 7/4/5 |
| | *baseline* vs. *alternative #2* | 4/3/9 |
| | *baseline* vs. *alternative #3* | 4/4/8 |
| | *baseline* vs. *alternative #4* | 10/1/5 |
| | *baseline* vs. *larger kernel* | 6/3/7 |
| MUC | *baseline* vs. *alternative #1* | 19/17/60 |
| | *baseline* vs. *alternative #2* | 22/10/64 |
| | *baseline* vs. *alternative #3* | 13/8/75 |
| | *baseline* vs. *alternative #4* | 51/9/36 |

Table 4: Results of weight/scale/kernel ablation studies. Numbers indicate datasets where the first feature significantly outperformed (win), underperformed (loss), or showed no significant difference (non-significant) compared to the second feature ($p < 0.05$).

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise. A low $p$-value ($p < 0.05$) indicates the difference between correlations is statistically significant, and if significant, the feature with the higher absolute correlation is declared the "winner".

We carried out a preliminary small-scale ablation study on weights and scales for MSG/MUC algorithms. Our baseline used scales $[1, 2, 4, 8]$ with weights $[0.4, 0.3, 0.2, 0.1]$ (*baseline*). We compared this against three alternatives: (1) the same scales with reversed weights $[0.1, 0.2, 0.3, 0.4]$ (*alternative #1*); (2) fewer scales $[1, 4]$ with weights $[0.6, 0.4]$ (*alternative #2*); (3) linearly incremental scales $[1, 2, 3, 4]$ with baseline weights (*alternative #3*) and (4) larger scales $[8, 16, 32, 64]$ with baseline weights (*alternative #4*). For MSG, we also evaluated using a larger $5 \times 5$ kernel size. For MSG, each comparison is carried out 16 times (number of datasets). For MUC, we carried out the comparison separately for each bit setting $\in \{3, 4, 5, 6, 7, 8\}$ and reported the total number of win/loss/non-significant ratio ($16 \times 6 = 96$ runs).

| Comparison | Result (win/loss/non-significant) |
|---|---|
| Colored MSG vs. Grayscale MSG | 6/2/8 |
| Multi-scale MSG vs. Single-scale MSG | 8/2/6 |
| Multi-scale MUC vs. Single-scale MUC | 5/1/10 |
| MSG vs. Edge Density | 5/2/9 |
| MSG vs. Patch-symmetry (Kyle-Davidson et al., 2023) | 7/1/8 |

Table 5: Results of feature comparison ablation studies. Numbers indicate datasets where the first feature significantly outperformed (win), underperformed (loss), or showed no significant difference (non-significant) compared to the second feature ($p < 0.05$).

Based on our ablation study, we selected the *baseline* weight and scale configuration for MSG and MUC algorithms to conduct subsequent experiments. While our ablation study guided the selection of a baseline configuration, we recognize that many other parameter combinations could be explored. A more heavily tuned setup might yield even better results, but we opted for a simpler baseline to focus on the broader methodological implications without spending excessive time on hyperparameter optimization.

We then extended our investigation by examining various algorithmic variations, including colored versus grayscale MSG, single-scale versus multi-scale implementations, and comparative analyses of different feature extraction techniques. These comparisons encompassed edge density and patch-based symmetry methods previously introduced by Davidson et al. Kyle-Davidson et al. (2023). Table 5 highlights the effectiveness of colored MSG, multi-scale techniques, and MSG over the baseline alternatives.

## A.6 CONCLUSION

In this work, we introduce novel visual features that significantly outperform those traditionally employed in visual complexity literature, enhancing the existing segmentation-based model while preserving its interpretability. Drawing inspiration from the familiarity bias, we explore a previously unexamined dimension of visual complexity: the element of surprise. We first developed the Surprising Visual Genome (SVG) dataset, created based on the well-studied Visual Genome collection, to systematically investigate semantic factors in complexity perception. Using this resource, we conducted human participant studies that revealed a consistent tendency to perceive unfamiliar, surprising, or bizarre images as visually more complex. We then used LLMs to automate the generation of surprise scores with explicit reasoning capabilities, creating a scalable approach to quantifying this subjective dimension of visual complexity without sacrificing interpretability. Our final model—which integrates low-level visual features, segmentation and class counts, and LLM-generated surprise scores—not only outperforms all previous handcrafted feature approaches but also achieves competitive results against the best-supervised model while maintaining interpretability.