

Is Your Image a Good Storyteller?

Xuijie Song¹, Xiaoyi Pang¹, Haifeng Tang², Mengyue Wu^{1*}, Kenny Q. Zhu^{3*}

¹ X-LANCE Lab, Department of Computer Science and Engineering

MoE Key Lab of Artificial Intelligence, AI Institute

Shanghai Jiao Tong University, Shanghai, China

²China Merchants Bank Credit Card Center, Shanghai, China

³University of Texas at Arlington, Arlington, Texas, USA

¹ {xiujiesong, fointpang, mengyuewu}@sjtu.edu.cn, ² thfeng@cmbchina.com, ³ kenny.zhu@uta.edu

Abstract

Quantifying image complexity at the entity level is straightforward, but the assessment of semantic complexity has been largely overlooked. In fact, there are differences in semantic complexity across images. Images with richer semantics can tell vivid and engaging stories and offer a wide range of application scenarios. For example, the Cookie Theft picture is such a kind of image and is widely used to assess human language and cognitive abilities due to its higher semantic complexity. Additionally, semantically rich images can benefit the development of vision models, as images with limited semantics are becoming less challenging for them. However, such images are scarce, highlighting the need for a greater number of them. For instance, there is a need for more images like Cookie Theft to cater to people from different cultural backgrounds and eras. Assessing semantic complexity requires human experts and empirical evidence. Automatic evaluation of how semantically rich an image will be the first step of mining or generating more images with rich semantics, and benefit human cognitive assessment, Artificial Intelligence, and various other applications. In response, we propose the Image Semantic Assessment (ISA) task to address this problem. We introduce the first ISA dataset and a novel method that leverages language to solve this vision problem. Experiments on our dataset demonstrate the effectiveness of our approach. Our data and code are available at: <https://github.com/xiujiesong/ISA>.

1 Introduction

How complex can a picture be? What kind of story can be told via a single picture? As the saying goes, “a picture is worth a thousand words”. However, not every picture contains such rich information. The Cookie Theft picture (Figure 1 (a)) is a good exemplar of complex semantic information expressed via visual language. It is a well-known picture commonly used to assess language and cognitive abilities in humans. It was first introduced in the Boston Diagnostic Aphasia Examination published in 1972 (Goodglass, Kaplan, and Weintraub 2001) and remains widely utilized to this day.

Many studies (Cummings 2019; Tasnim et al. 2022) have revealed the reasons behind the success of this picture. Its



Figure 1: The Cookie Theft pictures. (a) is the original version and (b) is an updated version.

essence is being a “good storyteller,” capable of telling a complete and engaging story. Based on the research of the psychologists, two of its most important characteristics can be summarized as follows: (1) It contains a rich but not excessive number of entities, making it well-suited for eliciting longer narrative descriptions. (2) It is rich in semantics, enabling it to tell an interesting story. The semantics are derived from reasonings made by observing the entities and their relationships in the image. For instance, the Cookie Theft tells a story about two children attempting to steal cookies from a jar when their mother is not looking. The mother-child relationship between the characters in the image is deduced through further reasoning based on observing the content in the image.

Though the Cookie Theft picture is widely used, there are still limitations. It is outdated since it has been proposed for half a century and it cannot be well applied to different cultures (Berube et al. 2019; Steinberg, Lyden, and Davis 2022). To avoid these issues, people often have to modify or replace the image in different application scenarios (Berube et al. 2019; Hussein et al. 2015; Domínguez et al. 2006; Oh et al. 2012; Prasad, Dash, and Kumar 2012). For instance, Figure 1 (b) is an updated version of the Cookie Theft. This means more images of this kind are necessary.

Besides, this kind of image is not only useful for humans but also for Artificial Intelligence (AI). With the development of vision models, especially Large Vision-Language Models (LVLMs), their abilities are increasing rapidly. Simple images with less semantics are not challenging enough for them to understand or generate anymore, so more im-

*Corresponding authors.

ages with rich semantics will definitely be beneficial for both training and evaluation (Song et al. 2024).

The internet or existing image datasets contain a lot of images, including some high-quality images that we expect. However, due to their scarcity, identifying and locating these high-quality images amidst the vast array of webly images can be a daunting task. Therefore, efficient methods for scoring and selecting these images are crucial. Furthermore, with the advancement of image generation models (Rombach et al. 2021; Ramesh et al. 2021, 2022), they are also increasingly capable of helping us generate more images. Thus, automatic semantic complexity assessment can also be used to assess the semantic complexity of generated images. Generally, it is the necessary path for obtaining semantically rich images.

Currently, though there are some research works about Image Assessment, like Image Quality Assessment (IQA) (Fang et al. 2020; Ying et al. 2020), Image Aesthetics Assessment (IAA) (He et al. 2022; Yi et al. 2023), and Image Complexity Assessment (ICA) (Saraee, Jalal, and Betke 2020; Feng et al. 2023), no one focuses on assessing the semantic complexity of images. In order to fill this research blank, we propose the **Image Semantic Assessment (ISA)** task to assess the semantic complexity of images.

Considering entities are the foundation of semantics and the complexity requirements for these two aspects may vary in different application scenarios, ISA task assesses images from both two levels: 1) At the *entity* level, we assess the entity richness of images, similar to the idea of ICA task (Feng et al. 2023), which we refer to as the Entity Complexity Scoring task; 2) At the *semantic* level, we propose the Semantic Complexity Scoring task to assess the higher-level semantic complexity of images. Note that this sub-task is the core of our proposed ISA task.

To promote the research on ISA task, we built the first ISA dataset with 2,946 images. Each image is annotated with the two corresponding scores by three annotators. Besides, a corresponding method called **Vision-Language collaborative ISA method (VLISA)** is proposed for this novel task. It first uses a Large Vision-Language Model (LVLM), such as GPT-4o (OpenAI 2023), as a feature extractor to extract semantic information in natural language form from images. Then, a regression model is trained to predict the score of images. Our contributions are as follows:

1. As far as we know, we are the first to propose the ISA task, which aims to automatically assess semantic complexity in an image. It can be used to identify high-quality images with rich semantics and evaluate image generation models, etc.

2. We construct the first ISA dataset, consisting of 2,946 images and human scores, that supports the ISA task. Our dataset includes images of varying semantic complexity, which helps models learn the ability to assess semantic complexity.

3. To effectively assess the semantic complexity of images, we propose a simple yet effective method that collaboratively utilizes language and visual information. Experiments show that ISA task is challenging for traditional vision models like ViT (Dosovitskiy et al. 2021) and our pro-

posed method significantly outperforms other baseline models on the Semantic Complexity Scoring task.

2 ISA Dataset Construction

In this section, we introduce our ISA data collection and annotation process, as well as the related data analysis.

2.1 Data Collection

We collected our images from Pinterest¹. After collecting images, we filtered out duplicated images using imagededup². To ensure high quality, we also manually excluded low-quality images that were blurry, watermarked or contained unnecessary text. After filtering, we finally retained 2,946 images in our dataset.



Figure 2: Samples from the proposed ISA dataset. ES and SS stand for Entity Score and Semantic Score respectively.

2.2 Data Annotation

For each image, we annotate it with two scores: an Entity Score and a Semantic Score. They correspond to the Entity Complexity Scoring task and the Semantic Complexity Scoring task, respectively. For each score, the images are first annotated on a scale from 1 (Low) to 5 (High). Then, these scores are normalized to the [0,1] range (Feng et al. 2023), and the average of these normalized scores is calculated as the final score.

¹<https://www.pinterest.com/>

²<https://github.com/idealo/imagededup>

Entity Score The scoring criteria (Figure 3) for the Entity Score are based on the richness of entities in the image. Entity Score differs slightly from that of the ICA task (Feng et al. 2023): we emphasize the richness of entities, e.g., an image with only complex lines would not receive a high score. Note that, since we do not always expect images to be overly cluttered and overwhelming, a higher Entity Score does not necessarily indicate a better image.

Score	Entity
1	The image contains very few entities, resulting in a simple composition.
2	The image contains more entities, but the overall number is still limited.
3	The image contains a rich variety of entities, with a full composition that does not appear visually crowded. Refer to the Cookie Theft picture.
4	The image contains an even richer variety of entities, with the content appearing somewhat visually crowded.
5	The image contains a very large and diverse number of entities, resulting in a very crowded composition.

Figure 3: Annotation criteria of Entity Score. The referenced Cookie Theft refers to the updated version.

Semantic Score To help annotators understand semantic complexity, we construct detailed annotation criteria of Semantic Score (Figure 4). The scoring criteria consist of five dimensions: event, connection between events, visual clues, storytelling, and interest level of the story. The connection between events primarily refers to their causal relationships. For instance, because “there’s a broken bowl on the floor,” (\rightarrow) “the mother is spanking the boy.” The definition of visual clues is entities capable of inferring new semantic conclusions. Depending on the type of conclusions inferred, visual clues encompass different categories: time, location, characters, character relationships, events, event relationships, character mental states, etc (Song et al. 2024). For example, a “Christmas tree” suggests the Christmas season, serving as a clue for time reasoning.

Specifically, the interpretation of Figure 4 is as follows: (1). If the image does not depict any event, or if it features only a very limited variety of events (e.g., running, swimming, etc.), assign it a score of 1. (2). If there are some kinds of events in the image, but the events are unrelated and there are almost no clues to infer additional information, give it a score of 2. (3). Assign a score of 2 when there are some events in the image with a slight connection between them or a few clues that suggest additional information, but the image does not convey a clear story or the story’s appeal is minimal. (4). For images rated 3 or above, there must be connections between events and visual clues present in the image. (5). The differences between ratings of 3, 4, and 5 primarily lie in the richness and number of these connections and clues.

Annotation Process Each image is annotated by three annotators, each assigning the two scores. The annotators are mostly undergraduate or graduate students aged 20 to 25.

We begin by providing training to the annotators. For each annotation score, several examples are provided for them to refer to. They are then asked to annotate a small sample set of images as a test first. Only annotators who pass the test are allowed to participate in the subsequent formal annotation.

During the annotation process, to ensure labeling quality, we maintain ongoing communication with the annotators, conduct regular spot checks on annotated image groups, and provide prompt feedback on any inaccurate annotations. If a sample from a particular group has poor annotation results, we will discard the labels for that group and re-label the group of images. We also provide immediate assistance if they encounter any issues during the annotation process.

2.3 Dataset Analysis

Annotation Consistency In line with established standards (Kong et al. 2016; Ying et al. 2020; Feng et al. 2023), we assess the consistency between annotators by using the Pearson Correlation Coefficient (PCC), Spearman’s Rank Correlation Coefficient (SRCC), and Kendall’s tau correlation for each pair of annotations. For Entity Score, the average PCC, SRCC, and Kendall’s tau are 0.836, 0.827, and 0.762 respectively. The average PCC, SRCC, and Kendall’s tau of Semantic Score are 0.799, 0.798, and 0.729 respectively. This demonstrates the consistency of our data annotation. In addition, following the crowdsourcing assessment studies conducted for IQA, IAA, and ICA (Hosu et al. 2020; Sahaan, Hanjalic, and Redi 2016; Feng et al. 2023), we compute the Intra-class Correlation Coefficient (ICC) for our annotations to measure the inter-rater reliability. The ICCs of Entity Score and Semantic Score are 0.937 and 0.922 respectively, which shows the reliability and consistency of our annotation.

Dataset Case Analysis Figure 2 shows some samples of our dataset. We can see that images with more entities are scored with higher Entity Scores, and images with more visual clues and telling more engaging stories are scored with higher Semantic Scores. We can also see that the relationship between Entity Score and Semantic Score is not entirely positively correlated. Even though some images contain few entities, for example, Figure 2 (e), they can still tell an interesting story. Images with a variety of entities can also contain little semantic information, for instance, Figure 2 (d).

Dataset Statistics Table 1 shows the distribution of our dataset. We randomly split the data into a training set, a validation set and a test set in a 6:2:2 ratio.

Score	[0, 0.2)	[0.2, 0.4)	[0.4, 0.6)	[0.6, 0.8)	[0.8, 1]
ES	476	789	999	294	388
SS	767	826	828	382	143

Table 1: Dataset label distribution. ES and SS stand for Entity Score and Semantic Score respectively.

Score	Event	Relationship between events	Visual clue	Storytelling	Interest level of the story
1 star 🤔	Few	None	None	The image does not tell a story.	--
2 stars 😐	Some	Few	Few	The image does not tell a complete story or makes it difficult to understand what is happening.	Boring
3 stars 😃	Some	Some	Some	The image tells a complete story.	Interesting
4 stars 😊	Some	Several	Several	The image tells a complete story.	Interesting
5 stars 😄	Abundant	Abundant	Abundant	The image tells a complete story.	Interesting

Figure 4: Annotation criteria of Semantic Score.

3 Method

In order to lay the foundation for the ISA task, we propose a novel baseline method to perform the task. Since we expect to assess images from a higher semantic level, and language can usually express semantics more directly than images, we believe hybrid utilization of both visual and language information will be helpful for ISA task. Thus, we propose the Vision-Language collaborative ISA (**VLISA**) method.



Figure 5: Pipeline of our proposed VLISA method.

As shown in Figure 5, VLISA has two components: a Feature Extractor and a Discriminator. This design follows the typical flow of IAA systems, which consists of a Feature Extraction phase and a Decision Phase (Deng, Loy, and Tang 2017). Specifically, we first use an LVLM as the Feature Extractor to extract semantic information in natural language form as features from images. We adopt GPT-4o (OpenAI 2023) as the default Feature Extractor in this paper, considering its strong capability. Then, we use a Discriminator model to rate the input image based on the extracted features, optionally including the image itself. One advantage of using LVLM as the Feature Extractor is that it can mitigate the impact of different image styles and types and focus more on semantics in images. Based on different feature extraction modes, we propose two versions of VLISA.

Naive VLISA The first way for GPT-4o to extract features from the input image is to make it describe the image in detail. The prompt is simple: `Describe this image in detail.` The generated description is later taken as the input of the Discriminator model.

Chain-of-Thought VLISA The second feature extraction method is inspired by Chain-of-Thought (CoT) (Wei et al. 2024). Referring to the annotation instruction of Cog-Bench (Song et al. 2024), we first ask GPT-4o to generate Chain-of-Reasonings (CoRs) from different aspects, then generate the description based on the CoRs. A CoR consists of visual clues and the conclusion drawn from the clues.

Specifically, we adopt seven categories of CoRs:

- **Special Time.** “Special time” refers to a time that requires observation of clues to deduce, rather than an obvious time like “daytime.”
- **Special Location.** “Special location” refers to a location that requires observation of clues to deduce, rather than an obvious location like “on the roadside.”
- **Character Role.** The roles or identities of characters in the images.
- **Character Relationship.** The relationships between characters in the images.
- **High-level Event.** “High-level events” refer to events that require observation of clues to deduce, rather than obvious actions like “running.”
- **Event Causal Relationship.** The causal relationships between events in the images.
- **Mental State.** Mental states of characters in the images.

Both the generated CoRs and description are later used as input for the Discriminator model.

4 Experiments

In this section, we present the experimental setup, results, and corresponding analysis.

Models For Vision models, we use ICNet (Feng et al. 2023) and ViT (Dosovitskiy et al. 2021) as our baseline models. ICNet is a model designed for the ICA task. ViT is a classic vision model. For VLISA, we use GPT-4o (OpenAI 2023) to extract features from the image and use ViLT (Kim, Son, and Kim 2021), BERT (Devlin et al. 2019), and Longformer (Beltagy, Peters, and Cohan 2020) as the Discriminator models. ViLT accepts both an image and its text features as input, while BERT and Longformer only accept text features as input.

Model	RMSE ↓	RMAE ↓	PCC ↑	SRCC ↑
ICNet	0.102 (0.001)	0.281 (0.0003)	0.918 (0.002)	0.909 (0.002)
ViT	0.094 (0.002)	0.271 (0.002)	0.929 (0.002)	0.923 (0.003)
Naive VLISA (ViLT)	0.079 (0.002)	0.249 (0.002)	0.952 (0.002)	0.949 (0.002)
Naive VLISA (BERT)	0.095 (0.002)	0.269 (0.002)	0.928 (0.001)	0.925 (0.002)
Naive VLISA (Longformer)	0.094 (0.0005)	0.268 (0.0005)	0.931 (0.0005)	0.928 (0.0005)
CoT VLISA (ViLT)	0.080 (0.002)	0.249 (0.002)	0.951 (0.002)	0.947 (0.003)
CoT VLISA (BERT)	0.111 (0.001)	0.288 (0.003)	0.902 (0.003)	0.897 (0.003)
CoT VLISA (Longformer)	0.111 (0.001)	0.287 (0.001)	0.901 (0.002)	0.893 (0.001)

Table 2: Performance of different methods on the Entity Complexity Scoring task. ↑ indicates that the larger the value, the better. ↓ indicates that the smaller the value, the better.

Model	RMSE ↓	RMAE ↓	PCC ↑	SRCC ↑
ICNet	0.191 (0.001)	0.389 (0.002)	0.634 (0.005)	0.626 (0.005)
ViT	0.182 (0.001)	0.383 (0.001)	0.677 (0.003)	0.667 (0.004)
Naive VLISA (ViLT)	0.148 (0.001)	0.339 (0.002)	0.799 (0.004)	0.791 (0.005)
Naive VLISA (BERT)	0.149 (0.001)	0.337 (0.002)	0.798 (0.003)	0.785 (0.003)
Naive VLISA (Longformer)	0.148 (0.001)	0.332 (0.001)	0.800 (0.003)	0.789 (0.004)
CoT VLISA (ViLT)	0.140 (0.0005)	0.328 (0.001)	0.823 (0.002)	0.817 (0.003)
CoT VLISA (BERT)	0.144 (0.0005)	0.328 (0.001)	0.812 (0.001)	0.802 (0.002)
CoT VLISA (Longformer)	0.149 (0.002)	0.333 (0.002)	0.804 (0.005)	0.795 (0.007)

Table 3: Performance of different methods on the Semantic Complexity Scoring task.

Implementation Details We implement the models using PyTorch (Paszke et al. 2019) and Transformers (Wolf et al. 2020). Each model is trained and evaluated on either a single NVIDIA A10 GPU or a Tesla V100 GPU. For ICNet, we mainly follow the settings in the original paper. For ViT, ViLT, BERT, and Longformer, we train them with batch size 16. They are fine-tuned based on *vit-base-patch16-224*, *vilt-b32-mlm*, *bert-base-uncased*, and *longformer-base-4096*, respectively. The maximum text input length of ViLT and BERT is set to 512 tokens. Since the maximum text input length of pre-trained ViLT is 40, we randomly initialize its position embeddings. The maximum input length of Longformer is set to 1024 tokens. These models are trained, validated, and tested on our training, validation, and test sets, respectively. We repeat all experiments three times and calculate the mean and standard deviation.

Evaluation Metrics Following Feng et al. (2023), we use Root Mean Square Error (RMSE), Root Mean Absolute Error (RMAE), Pearson Correlation Coefficient (PCC), and Spearman’s Rank Correlation Coefficient (SRCC) as our evaluation metrics. The formulas for calculating RMSE and RMAE are

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

and

$$RMAE = \sqrt{\frac{1}{n} \sum_{i=1}^n |x_i - y_i|}, \quad (2)$$

where n is the sample size, x_i and y_i represent the i th label and predicted score.

The PCC is defined as

$$r_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3)$$

where \bar{x} and \bar{y} are the means of \mathbf{x} and \mathbf{y} .

The formula of SRCC is

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (4)$$

where $d_i = R(x_i) - R(y_i)$, $R(x_i)$ and $R(y_i)$ are the ranks of the i th image when labels and predicted scores are sorted in descending order.

Results Table 2 shows the results of the Entity Complexity Scoring task. Naive VLISA with a pre-trained language model (BERT/Longformer) as the Discriminator shows competitive performance compared to ViT. When both images and text features are input to the ViLT Discriminator, the model performs significantly better than ViT and other Naive VLISAs. Naive VLISA (BERT/Longformer) outperforms CoT VLISA (BERT/Longformer). One possible reason is that features extracted by Naive VLISA tend to focus more on describing the content at the entity level within the image. Conversely, the Feature Extractor in CoT VLISA extracts higher-level semantic information from the image, but it overlooks some entities. Naive VLISA (ViLT) is less affected by the type of text features, probably because the image itself is visible to it.

Table 3 shows the results of the Semantic Complexity Scoring task. We can see that predicting Semantic Score is more challenging than predicting Entity Score and traditional vision models cannot perform well on this task. Naive

Task	Model	RMSE ↓	RMAE ↓	PCC ↑	SRCC ↑
Entity	Naive VLISA (ViLT)	0.080 (0)	0.247 (0.0005)	0.951 (0.0005)	0.948 (0.0005)
	Naive VLISA (BERT)	0.096 (0.001)	0.268 (0.001)	0.927 (0.001)	0.929 (0.0005)
	Naive VLISA (Longformer)	0.100 (0.003)	0.276 (0.004)	0.924 (0.001)	0.924 (0.001)
Semantic	Naive VLISA (ViLT)	0.149 (0.001)	0.343 (0.0005)	0.799 (0.001)	0.789 (0.001)
	Naive VLISA (BERT)	0.158 (0.002)	0.349 (0.001)	0.77 (0.006)	0.757 (0.004)
	Naive VLISA (Longformer)	0.152 (0.001)	0.339 (0.0005)	0.790 (0.003)	0.780 (0.001)

Table 4: Performance of Naive VLISA with CogVLM2 as the Feature Extractor. Entity and Semantic refer to the Entity Complexity Scoring task and the Semantic Complexity Scoring task, respectively.

Model	CoT Feature	RMSE ↓	RMAE ↓	PCC ↑	SRCC ↑
CoT VLISA (ViLT)	CoRs + Description	0.140 (0.0005)	0.328 (0.001)	0.823 (0.002)	0.817 (0.003)
	w/o CoRs	0.143 (0.001)	0.335 (0.001)	0.818 (0.003)	0.808 (0.003)
	w/o Description	0.142 (0.002)	0.33 (0.004)	0.819 (0.007)	0.808 (0.008)
CoT VLISA (BERT)	CoRs + Description	0.144 (0.0005)	0.328 (0.001)	0.812 (0.001)	0.802 (0.002)
	w/o CoRs	0.145 (0)	0.332 (0.0005)	0.809 (0.001)	0.799 (0.001)
	w/o Description	0.148 (0.001)	0.332 (0.001)	0.801 (0.004)	0.792 (0.005)
CoT VLISA (Longformer)	CoRs + Description	0.149 (0.002)	0.333 (0.002)	0.804 (0.005)	0.795 (0.007)
	w/o CoRs	0.148 (0.001)	0.334 (0.002)	0.802 (0.003)	0.793 (0.001)
	w/o Description	0.149 (0.001)	0.331 (0.001)	0.800 (0.004)	0.791 (0.005)

Table 5: Ablation study of CoT VLISA on the Semantic Complexity Scoring task.

VLISAs show obviously better performance than ViT and ICNet. The possible reason is that with GPT-4o extracting semantic information from the images, the Discriminator in VLISA can perform score prediction at a higher semantic level. Consistent with the previous hypothesis, CoT VLISA shows better performance than Naive VLISA on this task. CoT VLISA (ViLT) shows the best performance. Comparing the performance of VLISAs and vision models on the two tasks highlights the importance of introducing the language modality for the Semantic Complexity Scoring task.

Generally speaking, Naive VLISA can perform well on both sub-tasks, and CoT VLISA can further improve the performance on the Semantic Complexity Scoring task.

Open-source LVLM as Feature Extractor To further validate the effectiveness of VLISA pipeline, we replace GPT-4o with an open-source LVLM, CogVLM2 (Hong et al. 2024), as the Feature Extractor in Naive VLISA. As shown in Table 4, we observe that using CogVLM2 as the Feature Extractor does not significantly degrade the models’ performance, especially for Naive VLISA (ViLT). This demonstrates the robustness of our approach.

Ablation Study CoRs and the description are the two main parts extracted by the Feature Extractor of CoT VLISA, so we validate their effectiveness in this section. Table 5 shows when either CoRs or the description are removed from the extracted text features, there may be a slight performance drop. Therefore, we recommend using both the CoRs and Description as input.

5 Analysis

When using VLISA to identify semantically rich images, we recommend starting with those that have high Seman-

tic Scores and then refining the selection based on the Entity Score to match the application scenario. Figure 6 shows some samples with scores predicted by Naive VLISA (Longformer). In the scenario of searching for images similar to the Cookie Theft picture, images with higher Semantic Scores and moderate Entity Scores are preferred, according to the design guidelines.

In Figure 6, we can see that images with the highest Semantic Scores generally tell more compelling stories, containing richer semantic information. Interestingly, without including the Cookie Theft image in the training set, the image with the second-highest Semantic Score is Cookie Theft (Figure 6 (b)). Based on Entity Score, we can further filter out images with too few or too many entities. For instance, (a), (e), and (g) contain too many entities, though they also tell interesting stories. That is, the remaining images above the orange line are more preferred by the Cookie Theft design principles. Note that although many images in our dataset are not real-world images, VLISA can still give appropriate Semantic Scores to these real-world images. For example, image (d) in Figure 6 has a high Semantic Score and it is actually quite similar to the Cookie Theft semantically. This demonstrates that our method can avoid the influence of types or styles of images to some extent.

For images with the lowest Semantic Scores under the orange line, there is either a single action (Figure 6 (l)) or no event at all (Figure 6 (k, m, n, o)), which is also consistent with our annotation design.

6 Related Work

Image Quality Assessment Image Quality Assessment (IQA) is a task to assess the quality of images. It mainly concerns various types of distortions introduced in stages

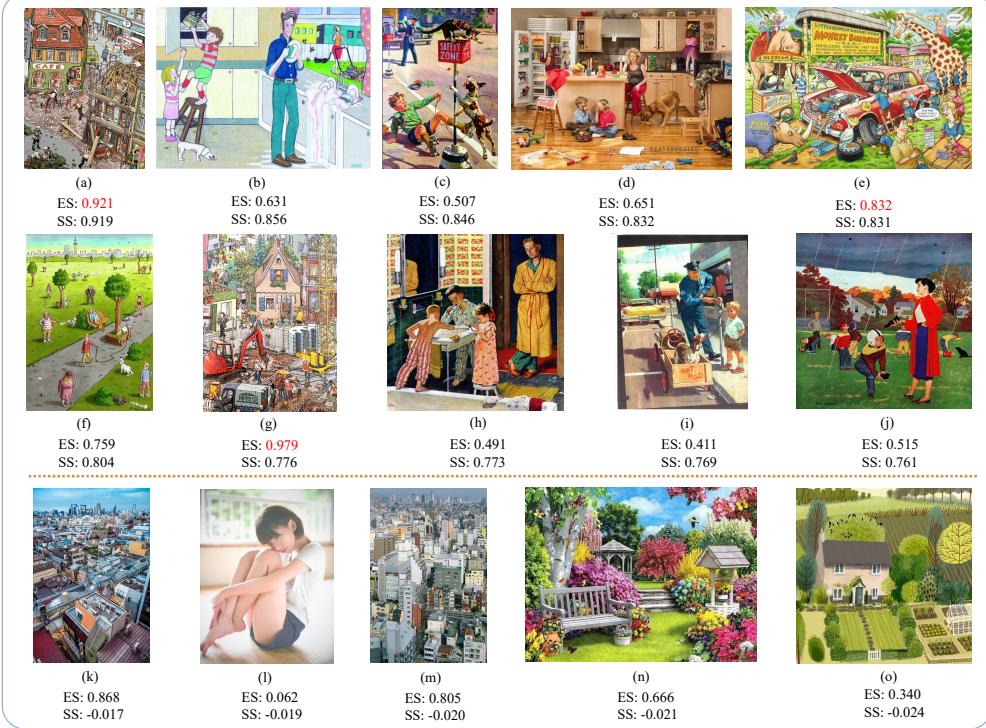


Figure 6: Case study. ES and SS stand for Entity Score and Semantic Score predicted by Naive VLISA (Longformer) respectively. The samples above the orange line are those with the highest Semantic Scores, sorted in descending order. The samples below the line are those with the lowest Semantic Scores. Red Entity Scores indicate that they are too high.

of the visual communication systems. The rapid growth of visual media has driven the development of many IQA methods (Zhai and Min 2020). Some IQA datasets including TID2013 (Ponomarenko et al. 2013), KonIQ-10k (Hosu et al. 2020), SPAQ (Fang et al. 2020) and PaQ-2-PiQ (Ying et al. 2020) etc. are proposed. With the development of LVLMs, Wu et al. (2024) propose Q-Bench to assess their abilities on low-level visual perception and understanding, which plays significant roles in IQA. The difference between IQA and our ISA task is that ISA task focuses on analyzing the semantic content (Zhang, Zhu, and Hwang 2015) of an image rather than its quality.

Image Aesthetics Assessment Different from IQA, Image Aesthetics Assessment (IAA) task assesses the aesthetics of an image from the perspective of its content. Typical IAA task seeks to computationally assess the quality of photos based on photographic rules (Deng, Loy, and Tang 2017). Several IAA datasets are proposed, for example, the Photo.net dataset (Joshi et al. 2011), the DPChallenge dataset (Datta, Li, and Wang 2008), and the TAD66K dataset (He et al. 2022) etc. As the development of image style transfer and AI painting, Artistic Image Aesthetic Assessment (AIAA) task is proposed to automatically evaluate artwork aesthetics (Amirshahi et al. 2015; Fekete et al. 2022; Yi et al. 2023). The difference between IAA and ISA task is that ISA assesses images based on their semantic richness.

Image Complexity Assessment Image Complexity Assessment (ICA) is proposed to assess the intricacy contained within an image (Forsythe 2009). It measures the richness of details and diversity within the image (Snodgrass and Vanderwart 1980). The SAVOIAS dataset (Saraee, Jalal, and Betke 2020) contains over 1,000 images and labels for IC analysis. Feng et al. (2023) built the first large-scale IC dataset with 9,600 annotated images IC9600 dataset and proposed a baseline model called ICNet. Compared to IQA and IAA, ICA task is more relevant to ISA task. Despite the differences, the Entity Richness Scoring in ISA shares similarities with the ICA task. However, the key distinction lies in ISA’s emphasis on a higher semantic level (Li et al. 2015), rather than merely evaluating complexity at the entity level.

7 Conclusion

In this paper, we propose a novel ISA task to identify storytelling images with rich semantics. We propose the first ISA dataset consisting of an Entity Complexity Scoring task and a Semantic Complexity Scoring task. We also propose a simple yet effective method called VLISA as our baseline model for this task. We believe this task will have a wide range of applications in the future. For example, with the Entity Score and Semantic Score, images with different semantic complexity levels can be selected. It can also facilitate AI models in understanding and generating images with richer semantics.

Ethical Statement

We follow the Terms of Service of Pinterest to collect the images in our ISA dataset. Our dataset will be available to download for research purposes only, which is in compliance with the Terms of Service of Pinterest.

Acknowledgments

Mengyue Wu was supported by National Natural Science Foundation of China (Grant No.92048205), the CMB Credit Card Center-SJTU joint research grant, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Jiangsu Technology Project (No.BE2022059-2) and Guangxi major science and technology project (No. AA23062062). Kenny Q. Zhu was partially supported by National Science Foundation (NSF) Award No. 2349713.

References

- Amirshahi, S. A.; Hayn-Leichsenring, G. U.; Denzler, J.; and Redies, C. 2015. JenAesthetics Subjective Dataset: Analyzing Paintings by Subjective Scores. In Agapito, L.; Bronstein, M. M.; and Rother, C., eds., *Computer Vision - ECCV 2014 Workshops*, 3–19. Cham: Springer International Publishing. ISBN 978-3-319-16178-5.
- Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150*.
- Berube, S.; Nonnemacher, J.; Demsky, C.; Glenn, S.; Saxena, S.; Wright, A.; Tippett, D. C.; and Hillis, A. E. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American journal of speech-language pathology*, 28(1S): 321–329.
- Cummings, L. 2019. Describing the cookie theft picture: Sources of breakdown in Alzheimer’s dementia. *Pragmatics and Society*, 10(2): 153–176.
- Datta, R.; Li, J.; and Wang, J. Z. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE International Conference on Image Processing*, 105–108.
- Deng, Y.; Loy, C. C.; and Tang, X. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4): 80–106.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Domínguez, R.; Vila, J. F.; Augustovski, F.; Irazola, V.; Castillo, P. R.; Escalante, R. R.; Brott, T. G.; and Meschia, J. F. 2006. Spanish Cross-Cultural Adaptation and Validation of the National Institutes of Health Stroke Scale. *Mayo Clinic Proceedings*, 81(4): 476–480.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual Quality Assessment of Smartphone Photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Fekete, A.; Pelowski, M.; Specker, E.; Brieber, D.; Rosenberg, R.; and Leder, H. 2022. The Vienna Art Picture System (VAPS): A data set of 999 paintings and subjective ratings for art and aesthetics research. *Psychology of Aesthetics, Creativity, and the Arts*.
- Feng, T.; Zhai, Y.; Yang, J.; Liang, J.; Fan, D.-P.; Zhang, J.; Shao, L.; and Tao, D. 2023. IC9600: A Benchmark Dataset for Automatic Image Complexity Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8577–8593.
- Forsythe, A. 2009. Visual Complexity: Is That All There Is? In Harris, D., ed., *Engineering Psychology and Cognitive Ergonomics*, 158–166. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-02728-4.
- Goodglass, H.; Kaplan, E.; and Weintraub, S. 2001. *BDAE: The Boston Diagnostic Aphasia Examination*. Philadelphia, PA: Lippincott Williams & Wilkins.
- He, S.; Zhang, Y.; Xie, R.; Jiang, D.; and Ming, A. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 942–948. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. 2024. CogVLM2: Visual Language Models for Image and Video Understanding. *arXiv preprint arXiv:2408.16500*.
- Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.
- Hussein, H. M.; Abdel Moneim, A.; Emara, T.; Abd-elhamid, Y. A.; Salem, H. H.; Abd-Allah, F.; Farrag, M. A.; Tork, M. A.; Shalash, A. S.; Ezz el dein, K. H.; Osman, G.; Georgy, S. S.; Ghali, P. G.; Lyden, P. D.; and Moustafa, R. R. 2015. Arabic cross cultural adaptation and validation of the National Institutes of Health Stroke Scale. *Journal of the Neurological Sciences*, 357(1): 152–156.
- Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and Emotions in Images. *IEEE Signal Processing Magazine*, 28(5): 94–115.
- Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*.

- Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14, 662–679. Springer.
- Li, P.; Wang, H.; Zhu, K. Q.; Wang, Z.; Hu, X.; and Wu, X. 2015. A large probabilistic semantic network based approach to compute term similarity. *IEEE Transactions on Knowledge and Data Engineering*, 27(10): 2604–2617.
- Oh, M. S.; Yu, K. H.; Lee, J.-H.; Jung, S.; Ko, I. S.; Shin, J. H.; Cho, S.-J.; Choi, H.-C.; Kim, H. H.; and Lee, B. 2012. Validity and Reliability of a Korean Version of the National Institutes of Health Stroke Scale. *Journal of Clinical Neurology (Seoul, Korea)*, 8: 177 – 183.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Red Hook, NY, USA: Curran Associates Inc.
- Ponomarenko, N.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Jin, L.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; and Kuo, C.-C. J. 2013. Color image database TID2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing (EUVIP)*, 106–111.
- Prasad, K.; Dash, D.; and Kumar, A. 2012. Validation of the Hindi version of National Institute of Health Stroke Scale. *Neurology India*, 60 1: 40–4.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8821–8831. PMLR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.
- Saraee, E.; Jalal, M.; and Betke, M. 2020. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195: 102949.
- Siahaan, E.; Hanjalic, A.; and Redi, J. 2016. A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal. *IEEE Transactions on Multimedia*, 18(7): 1338–1350.
- Snodgrass, J. G.; and Vanderwart, M. L. 1980. 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.
- Song, X.; Wu, M.; Zhu, K. Q.; Zhang, C.; and Chen, Y. 2024. A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision-Language Models. *arXiv preprint arXiv:2402.18409*.
- Steinberg, A.; Lyden, P. D.; and Davis, A. P. 2022. Bias in Stroke Evaluation: Rethinking the Cookie Theft Picture. *Stroke*, 53(6): 2123–2125.
- Tasnim, M.; Eghaghi, M.; Diep, B.; and Novikova, J. 2022. DEPAC: a Corpus for Depression and Anxiety Detection from Speech. In Zirikly, A.; Atzil-Slonim, D.; Liakata, M.; Bedrick, S.; Desmet, B.; Ireland, M.; Lee, A.; MacAvaney, S.; Purver, M.; Resnik, R.; and Yates, A., eds., *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 1–16. Seattle, USA: Association for Computational Linguistics.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.
- Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; and Lin, W. 2024. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *ICLR*.
- Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; and Rosin, P. L. 2023. Towards Artistic Image Aesthetics Assessment: A Large-Scale Dataset and a New Method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22388–22397.
- Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3575–3585.
- Zhai, G.; and Min, X. 2020. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63: 1–52.
- Zhang, K.; Zhu, K.; and Hwang, S.-w. 2015. An association network for computing semantic relatedness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.