

# Is Relevancy Everything?

## A Deep Learning Approach to Understand the Coupling of Image and Text

Jingcun Cao

The University of Hong Kong, jcao@hku.hk

Xiaolin Li

London School of Economics and Political Science, x.li166@lse.ac.uk

Lingling Zhang

China Europe International Business School, lzhang@ceibs.edu

### Abstract

Firms increasingly use a combination of image and text when displaying products or engaging consumers. Existing research has examined the role of text and image in consumer choice separately, without systematically considering the semantic relationship between them. In this research, we examine the effect of image- and text-based product representation and explore how image-text congruence affects consumer preference. We propose a state-of-the-art deep learning model to measure information congruence between products cover image and text description. Our Two-Branch Neural Networks model incorporates Wide-ResNet-50-2 (WRN) and BERT (Bidirectional Encoder Representations from Transformers) to capture the semantic relationship between image and text. Using individual-level consumption data from an online reading platform, we further examine the impacts of the image-text congruence on consumer behavior and identify a U-shape relationship: consumers prefer a product when the image-text congruence is either high or low, but not in the middle level. We explore the underlying mechanisms using an online study and find that the effect of high congruence is driven by information fluency, while that of low congruence is due to a “surprise-evoked” information elaboration. Our study contributes to the literature of consumer information processing, and provides important managerial implications to marketing practitioners and policy makers.

*Keywords:* multi-media formats; image-text congruence; visual analytics; consumer information processing; deep learning

# 1 Introduction

In the digital era, firms increasingly utilize image and text together to present their products and to communicate with consumers. For example, Netflix presents images and short descriptions for its entertainment products. On social media platforms such as Twitter and Facebook, brands increasingly include images along with their posts. In practice, firms often pair images with their textual content just based on heuristic considerations. Limited research exists on how to couple image and text and how consumers would respond to the different combinations of them. How do we measure the semantic relationship between image and text? What type(s) of image and text combinations can yield the most positive response from consumers? In this paper, we set out to answer these questions.

Marketing research has long been interested in how unstructured product information such as image and text affects consumer choice and perception. For text data, marketing researchers have used various methods including entity extraction, topic extraction and relation extraction to study how the sentiment and content of text can affect consumer preferences (Berger et al. 2020). For images, research has shown that including image(s), especially high-quality ones, in product communications can have a positive effect on consumer choice (Shedler and Manis 1986, Zhang et al. 2021b, Zhang and Luo 2018). Several important image features have been identified to affect consumer information processing, including aesthetic features such as colorfulness, color composition, and surface size as well as content features such as the presence of human face in the image (Finn 1988, Li and Xie 2020, Pieters and Wedel 2004, Wedel and Pieters 2015). While extant image analysis research has largely relied on manual coding to extract features, recent developments in machine learning allow researchers to study images in a more automatic way. For example, Zhang et al. (2021b) use machine learning to conduct large-scale image analytics and examine how dwelling images affect demand on Airbnb.

One significant limitation of the existing research, however, is that the multiple information formats have almost always been examined independently. While many empirical settings involve both images and text in product communications, the semantic relationship between them has not yet been sufficiently studied. Must the information contained in the image be consistent with the information conveyed in the text? What happens if the two formats are not aligned in the information conveyed? Answers to these questions have important managerial implications and can

shed light on how managers can optimally pair images and text to boost consumer preferences.

The image and text relationship has been studied by consumer behavior researchers. Using lab experiments, Heckler and Childers (1992) and Lee and Mason (1999) examined picture relevance in print ads and found that advertisements involving relevant and unexpected images tend to be more effective than those with irrelevant or expected images. In a more recent paper, Li and Xie (2020) conducted a large-scale image analysis and found that picture relevance could affect engagement with user generated content (UGC) on Twitter and Instagram. Findings from Li and Xie (2020) confirm that it is important for marketers to understand the relationship between images and text, but also call for future research that conducts the analysis without relying on human coding of images. In this study, we propose an innovative method to evaluate image and text in a more systematic and scalable way. Our method employs cutting-edge image and text embeddings, which enable us to extract the semantic meaning from images and text, respectively. Furthermore, we jointly train the image and text embeddings by fitting a supervised deep learning model with two-branch neural networks to measure the image-text congruence. By jointly analyzing these two information sources, our method captures the semantic relationship between them better than analyzing each source separately. A model comparison analysis confirms that our proposed method has superior performance compared to status quo methods.

We apply this two-branch deep learning model to the data from a platform that offers ebooks and audio-books to young readers. As a leading player in the market, the collaborating platform provides extracurricular reading materials to children between 5 and 16 years old. Our data include the detailed browsing and choice behaviors from more than 11,000 users over a period of seven months in 2019. We demonstrate that our proposed deep learning model can evaluate the semantic congruence between book profile images and book summaries (text) in an efficient and scalable way.

With the measured image-text congruence, we further examine its effect on product choice and explore potential mechanisms. We find that the image-text congruence has a U-shape relationship with consumer preference. Consumers tend to be more likely to choose a book when the congruence between its cover image and its text summary is either high or low. Interestingly, there is a dull zone around the middle level of congruence, suggesting that consumer preference is lowest when the book cover and summary have a medium level of congruence. This result is after controlling for the aesthetic features of images (colorfulness and color contrast) as well as the content of the

cover and book summary. We also find that the image-text congruence effect varies between ebooks and audio-books and fades over time. To understand the underlying mechanisms, we conduct an online lab study in which we specifically measure the *relevancy* and *expectancy* between each pair of cover image and text. Results confirm that high *relevancy* and low *expectancy* both can increase consumer preference. When *relevancy* is the driving force, consumers are likely affected by the fluency between information sources and make a quick choice. In contrast, when low *expectancy* is the driving factor, consumers are surprised about the lack of congruence between image and text and tend to spend more time processing the information, which can also lead to a higher choice likelihood. These findings advance researchers' and practitioners' understanding of how consumers respond to image and text congruence. Such an understanding can also help managers couple images and text descriptions to more effectively engage customers.

This study contributes to the marketing literature on product information communication in the following ways. First, we propose an innovative method to measure the congruence of image and text description. Our method jointly analyzes the semantic relationship between two very different unstructured data formats, namely image and text. The essence of this model can be generalized to more than two formats and beyond image and text. This method can also help researchers address other empirical questions such as how to detect misinformation and how to engage consumers through content marketing. Second, we identify an interesting U-shape relationship between image-text congruence and consumer preference. This pattern has not yet been documented in the literature, but we believe it has important implications due to the ubiquitous adoption of images in product communication and social media.

The paper proceeds as follows. We briefly discuss relevant literature in Section 2. In Section 3, we present the details of our proposed deep learning method on measuring image and text congruence, and compare its performance to benchmark models. In Section 4, we describe the empirical setting and the model. We discuss the results and present the online study to explore the mechanism in Section 5. In Section 6, we discuss the theoretical contributions, managerial implications, and future research directions.

## 2 Related Literature

Extant marketing research has been relatively silent on the role of media format on consumer information processing. The lack of research lags industry practice as firms increasingly employ

multiple formats to present their products, such as using image and text combinations. There is a nascent but rising stream of literature that explicitly examines these unstructured information sources in marketing applications. In this section, we briefly summarize the state of research on this topic and discuss how our paper contributes to the literature.

## **2.1 Text Analysis on Product Choice**

In marketing applications, the common techniques for text analysis fall into three categories: entity extraction, topic extraction, and relation extraction (Berger et al. 2020). Entity extraction involves identifying meaningful keywords and phrases from unstructured textual data, which are then used to generate psychological states or to predict consumer choice or market trends. For example, Berger and Milkman (2012) conducted sentiment analysis on newspaper articles and found that content that evokes negative or high-arousal emotions is more likely to be shared.

Entity extraction, however, suffers from a limitation that the semantic relationship among words or phrases is not captured in an automatic way. Topic extraction solves this problem and has thus been widely adopted in recent marketing applications. Common topic extraction methods include Latent Dirichlet Allocation (LDA) (Blei et al. 2003), which utilizes the co-occurrence pattern of words and phrases to infer the relationship among them. In marketing, Liu et al. (2016) demonstrated how topics extracted from social media platforms are used for market trend forecasting.

The third type of method, relation extraction, includes word embedding such as Word2Vec (Mikolov et al. 2013). This group of techniques uses deep learning models to reconstruct word representations via a vector space so that the distance between word vectors reflects the semantic (dis)similarities between them. Compared with topic modeling, word embedding methods take into consideration the sequence of word occurrence within a semantic context and thus can infer the semantic meaning of words. Timoshenko and Hauser (2019) used word embedding to identify customer needs from online user-generated content. They showed that the insights extracted through large-scale automatic text mining are comparable to those identified through traditional qualitative marketing research.

## **2.2 Image Attributes and Analysis**

An evolving stream of marketing literature is devoted to examining how different image attributes affect consumer perception and choice. The attributes studied belong to three major categories: image aesthetics (such as color composition), image sentiment, and the image content. Finn (1988)

found that picture colorfulness can increase the readership of print ads. Wedel and Pieters (2015) showed that the color composition of the central object plays a key role in projecting the gist of ad perception. Li and Xie (2020) examined user engagement on social network platforms and found that colorfulness, picture quality, and the presence of human faces can induce more engagement and sharing.

While these studies used human coders to extract pre-defined attributes, more recent studies developed machine learning methods to automatically identify image features. Using data from Airbnb, Zhang et al. (2021b) and Zhang et al. (2021a) used machine learning to identify what lower-level image attributes contribute to the perceived image quality and thus are related with demand. They identified systematic differences between verified and unverified images pertaining to several human-interpretable attributes and quantify the value of verified photos to be worth of thousands of dollars on average.

### **2.3 Interaction of Image and Text**

One important question remains: How do image and text jointly affect consumer perception and choice? As multi-media product presentations become the status quo on many online platforms, this question has important managerial implications.

Marketing researchers have long been interested in the interplay between text and visuals in marketing communication. One important concept in this domain is information congruence. In the context of print ads, Heckler and Childers (1992) proposed that congruence has two dimensions: *relevancy* and *expectancy*. *Relevancy* is defined as “material pertaining directly to the meaning of the theme.” The image and text in a print ad would be considered relevant if the information contained in the stimulus contributes to, rather than distracts from, the primary message being communicated. The second dimension, *expectancy*, refers to “the degree to which a piece of information falls into some predetermined pattern or structure evoked by the theme.” Using this framework, Heckler and Childers (1992) found that relevant and unexpected information can evoke elaborated processing and thus increase memory and recall. Lee and Mason (1999) further studied these two dimensions of congruence on consumer attitudes and found that relevant pictures in print ads lead to more favorable evaluations.

Despite the importance of the image-text interplay, empirical research is limited, partially because it is difficult to measure congruence in a scalable way. In particular, relevancy can be measured

more objectively than expectancy because the latter depends on consumers’ prior knowledge and conjectures. This explains the recent rise of empirical research on image-text relevancy. Using social media posts, Li and Xie (2020) concluded that image relevancy has a positive effect on user engagement on Twitter, but not on Instagram. We extend this stream of literature both methodologically and substantively. In terms of advancing the method, we propose an innovative deep learning model based on the two-branch Neural Networks to measure the image-text congruence. Our approach allows researchers and practitioners to evaluate holistically how much the information embedded in the image contributes to the theme contained in the text stimulus. With respect to the substantive contribution, our paper examines how image-text congruence affects consumer product choice and sheds light on the potential mechanisms.

It is worth mentioning that our research is different from the studies contrasting visual and textual stimuli. For example, Pieters and Wedel (2004) compared pictorial and text elements and found that the former captures consumer attention better than the latter in print ads. Zhang and Luo (2018) found that user-posted photos in online reviews have a positive impact on restaurant survival above and beyond the textual review content. Instead of evaluating which stimulus is more effective, our research focuses on understanding the semantic relationship between image and text, which can help managers boost interest for their products.

### **3 Image and Text Congruence: A Deep Learning Approach**

Marketers have recognized that the degree of congruence between products’ profile images and text descriptions can influence consumer decisions (Dew and Ansari 2018, Li and Xie 2020). However, measuring this image-text congruence has been challenging, especially if the goal is to capture the semantic similarity in an automatic and scalable manner. Image and text are two modalities of media that are difficult to compare directly. We first have to transform the two modalities into one common modality and then measure the congruence between them using this common modality. Drawing on recent developments in Computer Vision (CV) and Natural Language Processing (NLP), we propose a process that allows marketers to evaluate the cross-modalities semantic congruence between image and text. Our method employs the architecture of two-branch Neural Networks (Ge et al. 2021, Wang et al. 2018, Wei et al. 2020, Xu et al. 2020) and trains a deep learning model to measure the image-text congruence using three steps: (1) embedding the cross-modalities data (i.e., image and text) into respective numerical vectors, (2) computing the distance between the vectors

to measure the (dis)similarity, and (3) optimizing the loss function between predicted similarity and the manually labeled congruence data on the images and text based on back propagation. We conduct these three steps iteratively until the model converges. Although it is not the only method available to assess image and text congruence, we adopt this deep learning model based on two-branch Neural Networks because it captures the semantic meaning better than other existing methods. In Section 3.3, we compare our method with other benchmark methods, including both unsupervised and supervised approaches, and demonstrate that our approach exhibits the best performance on prediction accuracy and scalability. Next, we describe our approach in detail.

### 3.1 Image and Text Embeddings

To analyze images and text, existing methods typically involve extracting key information from the media to create “feature” variables. For example, for images, feature variables can be created based on the pixel value, the HSV (hue, saturation, and value) color, and the picture texture (Hardoon et al. 2004). For text, paragraphs can be tokenized into keywords and the relative frequency of keywords can be generated to create metrics such as “Term Frequency Inverse Document Frequency” (TF-IDF) (Ramos et al. 2003).

There are two limitations with the existing methods. First, the predefined feature extraction methods mainly rely on the statistical results of pixels (of image) or tokens (of text) instead of extracting real semantic meanings conveyed from the images and text. Second, the existing methods treat feature extraction and prediction as two isolated procedures, so that the extracted features may not fit the following prediction step and that the prediction results would not provide feedback to improve the extraction step. These limitations explain why the existing methods would perform poorly on semantic understanding tasks. In recent years, deep learning methods based on DNNs can process input data layer-wise to generate semantic embedding, which has delivered improved performance on various tasks in CV (He et al. 2016) and NLP (Devlin et al. 2018). Our proposed method adopts the state-of-the-art deep learning models of Deep Residual Networks (ResNet) and the Bidirectional Encoder Representation from Transformers (BERT) to extract the semantic meanings from images and text, respectively. We then connect the two-branch networks with a prediction network layer (transforming two modalities into a common modality), to jointly fine-tune the feature extraction and prediction on labeled annotations. This method connects two isolated procedures into a unified framework with better synergy and coordinates them for a more accurate congruence



measurement. With the trained model, images and text can be embedded into numeric vectors in a common modality, and then we calculate the distance between vectors to capture the semantic congruence between an image and text pair.

### 3.1.1 Image Embedding

For image processing, the popular underlying architecture is the Convolutional Neural Network (CNN) (Devlin et al. 2018, Krizhevsky et al. 2012), which uses 2D convolutional kernels to characterize the spatial relationship of the pixel matrix. It has been shown that rich high-level features can be captured by these kernels during the layer-by-layer convolution that gradually embeds the raw pixels into an abstract semantic space. ResNet (He et al. 2016) presents a milestone architecture in DNN, as it designs residual blocks to improve the effectiveness of very deep networks. Prior to ResNet, the performance of DNN can decrease when its depth increases. The residual blocks in ResNet reformulate the layers as learning residual functions with reference to the layer inputs instead of learning unreferenced functions. Because of these features, ResNet is one of the most widely used architectures in vision tasks (He et al. 2016, Zagoruyko and Komodakis 2016).

In this research, we adopt the Wide-ResNet-50-2 embedding (Zagoruyko and Komodakis 2016) pre-trained on ImageNet, which is an improved modification of the original ResNet method. Wide-ResNet (WRN) further proposes that it is more effective and efficient to widen the residual blocks with shallow depth. As a result, WRN outperforms ResNet on many vision tasks, including image classification and object detection. In this paper, the adopted WRN-50-2 is set with a depth of 50 and a widening factor of 2.

### 3.1.2 Text Embedding

For text, more advanced methods propose to generate an embedding vector for each word (or each text) by directly learning from the context rather than using pre-designed statistical algorithms such as TF-IDF. The core idea is to pretrain the network on a gigantic corpus such as Wikipedia and web pages so that universal semantic embeddings can be generated.

One of the current state-of-the-art methods is BERT (Devlin et al. 2018). BERT’s model architecture is a multi-layer bidirectional transformer encoder based on the original implementation in Vaswani et al. (2017). There are three types of embeddings in BERT—token embeddings, segment embeddings, and position embeddings, each of which is designed to capture token, segment, and position information, respectively. For a given WordPiece token, the input representation is con-

structed by summing the corresponding token, segment, and position embeddings (Wu et al. 2016). To learn the embeddings, BERT is pre-trained on the corpus with two unsupervised tasks. In the first task, “masked LM,” BERT randomly masks some percentage of the input tokens and then predicts those masked tokens. In the second task, “Next Sentence Prediction,” BERT constructs the “IsNext” sentence pairs and “NotNext” pairs randomly from the corpus to do binary identification. These two tasks help BERT learn embeddings by understanding the context and sentence relationships.

In this research, we adopt the BERT-base embedding pre-trained on Chinese Wikipedia (Cui et al. 2020). Our BERT-base has 12 layers (i.e., transformer blocks), 12 attention heads, and a hidden size of 768, with 110 million parameters in total.

### 3.2 Two-Branch Neural Networks for Image-Text Congruence

With image and text embedding, the next step is to jointly process them to understand the semantic congruence between the two media. To accomplish this, we need to represent the two modalities in a common modality (i.e., same vector dimension) (Hardoon et al. 2004). Deep learning models, with the advantage of a flexible end-to-end framework, can learn joint image-text embedding where multi-modal inputs are simultaneously processed for joint visual and textual understanding (Chen et al. 2019, Wang et al. 2018).

To do so, we adopt the popular architecture of two-branch neural networks (Ge et al. 2021, Wang et al. 2018, Wei et al. 2020, Xu et al. 2020). Figure 1 depicts our model architecture. The left CNN branch encodes the image RGB into image embeddings and the right Transformer (Trm) encodes text tokens  $(T_1, \dots, T_N)$  into text embeddings, where the image and text embeddings are specified to have the same length. Specifically, for the image embedding based on Wide-ResNet-50-2, we fix the parameters of all the layers except the last fully connected layer and reinitialize the vectors to a dimension of 768 so that it matches the output dimension of the BERT embeddings from the text. Since the last embedding layer of pre-trained Wide-ResNet-50-2 outputs 2,048 dimensional features, to align it with BERT, which outputs 768 dimensional features, we reinitialized the last fully connect layer of ResNet50 as a  $2048 \times 768$  matrix. More specifically, the embeddings of pre-trained Wide-ResNet-50-2 and BERT are  $M_1 \in R^{n \times 2048}$  and  $M_2 \in R^{n \times 768}$ , respectively. We set the batch size to be  $n$ , the re-initialized fully connected layer to be  $W \in R^{2048 \times 768}$ . All the parameters of Wide-ResNet-50-2 before this fully connect layer are frozen and untrainable. We only add this

linear transformation and after this layer, embeddings from the two branches have the same shape. Finally, we measure the congruence between the two branches based on cosine similarity as the model output. We note the averaged labeled congruence and model output as  $y_i$  and  $\hat{y}_i$ , where  $i \in \{1, \dots, n\}$ . The formalization of our model and loss function  $\ell$  can be written as

$$\hat{y} = \text{cosine}(M_1W, M_2)$$

$$\ell(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

---

Insert Figure 1 about here.

---

Next, we train the model with human labeled data on the image and text congruence. WRN and BERT are trained on hyperscale datasets such as ImageNet (Deng et al. 2009) and Wikipedia and thus have learned universal semantic embedding. As a result, they can be fine-tuned with a small amount of labeled data to accommodate the specific task at hand (Bengio 2012, Tan et al. 2018). We randomly sampled 2,000 pairs of color images and the corresponding texts, where the images are rescaled as 224×224 RGB images. For each pair, three trained independent annotators rated the level of congruence between the image and text (see Figure A1 in the Appendix for an illustration of the labeling interface). Three annotators are graduate students from major a research University in US. The generated labels (i.e., human annotations) are used as the ground truth. The correlation of the congruence ratings on the 2,000 pairs are reported in 1, where the high correlations between annotators suggest inter-annotator reliability of the congruence labeling. The model parameters are then estimated through supervised learning, by minimizing the Mean Square Error (MSE) loss between predictions and the ground-truth labels. For parameter tuning, the image-text pairs were split into a training set of 1,800 pairs and a test set of 200 pairs. We train the model for 100 epochs<sup>1</sup> with a batch size of 40 with a learning rate of  $1e^{-3}$ . The optimization process is presented in Figure 2, where the training loss is the MSE on each batch after five iterations and the test loss is the MSE on the test set after each epoch. The training loss shows that our model converges. We utilize the Stochastic Gradient Descent (SGD) as the optimizer to fine-tune the network. The model is implemented using PyTorch.

---

<sup>1</sup>Before we did the parameter tuning, we separated 200 samples from the training set as validation and observed that after about 80 epochs the validation performance reaches a stable optimal point. Therefore, we set the epoch as 100 using the whole training set.

---

Insert Figure 2 about here.

---

---

Insert Table 1 about here.

---

To evaluate the performance of our method, we compute the Pearson correlation between the model prediction and the human annotated label. The results are presented in Table 1. The correlation between the model prediction and the average of the three annotators is 0.741, which is positive and moderately strong. Therefore, it indicates that the trained model achieves satisfactory out-of-sample performance in terms of predicting human labels.

To further illustrate the congruence measure predicted by our model, we select three pairs from our sample, which correspond to low, medium, and high congruence between the image and text. For each pair, the image, text, and the model-predicted congruence score are presented in Table 2. As seen there, our model output demonstrates a reasonable level of agreement with human interpretation. For the first pair, the text does not seem to describe the image well and our model predicts the congruence to be low (0.140). For the pair in the middle, the text mentions the origin of life, which is somewhat related to the content of the image. Accordingly, the congruence predicted by our model is 0.523. Finally, the text and the image in the last pair are both related to the theme of “School of Elephants: A Trip to the Zombie Land,” which matches the high prediction (0.736) from our model.

---

Insert Table 2 about here.

---

### 3.3 Comparison with Benchmark Methods

In this section, we compare the performance of our proposed method to two common benchmark methods, both of which use established tools to encode images and text. Below, we demonstrate the superior performance of our method and describe how it differs from the benchmark methods.

To assess the semantic correlation between an image and text pair, our proposed method is to pull two modalities to a third common modality. Another possible approach is to transform one modality to the other; i.e. to “translate” an image into text or “translate” text into an image, with the former technically more realistic than the latter. Thus, we compare our method to this “images-to-text” method.

We adopt the Google Cloud Vision API (Bengio 2012) to extract tags from images and convert them into features. Using industry-leading machine learning techniques, Google Cloud Vision API detects a fraction of an image that matches a predefined tag and then assigns the tag to the image. For example, an image that contains a cat will be assigned a tag “cat” along with the associated probability. It also detects text (printed and handwritten) in an image using Optical Character Recognition (OCR). After “translating” images into text, we employ text processing techniques to measure the correlation between a pair of texts. We adopt LDA (Blei et al. 2003), which treats text as a bag of words and each word as a finite mixture over an underlying set of topics. Through LDA, each text is then represented as a distribution of topics. Prior to this topic modeling, we follow the common text pre-processing steps including creating word segmentation and removing stop words.

### 3.3.1 Benchmark Method 1: Unsupervised “Image-to-Text” Method

The first benchmark method uses an unsupervised approach to measure the semantic congruence between image content (converted into text tags through Google Cloud Vision API) and text topics (through LDA). Formally, for the  $i^{th}$  image and text pair  $(img_i, t_i)$ , let us denote the extracted labels of  $img_i$  as  $f_i$ , the topic number as  $k$ , and the corresponding topic vectors as  $(topic_i^f, topic_i^t)$ . Since the vector is a probability distribution on  $k$  topics, we have  $topic_* \in [0, 1]^k$  and  $\sum topic_* = 1$ . We treat the semantic contribution of each topic equally and compute the image-text congruence,  $r_i^u$ , as the cosine similarity between the two topic distributions:

$$r_i^u = \frac{topic_i^f * topic_i^t}{\|topic_i^f\| * \|topic_i^t\|}.$$

The cosine similarity computes the cosine value of the angle between two vectors. Larger cosine similarity values correspond to smaller angles between the topic vectors, indicating higher semantic similarity between the topics.

### 3.3.2 Benchmark Method 2: Supervised “Image-to-Text” Method

The unsupervised method ignores the fact that different topics may overlap semantically and that some topics may be more important than others in conveying the semantic meaning. To solve this problem, we build a supervised linear regression model trained on manual annotations to learn the coefficient of each topic. The dependent variable of this regression is the human annotation (i.e., whether or not the image and text is congruent). The independent variables are the differences between the text vectors for each topic,  $(topic_i^f - topic_i^t) \in [-1, 1]^k$ . This difference vector has  $k$

dimensions, each capturing the probabilistic difference along that topic. We then linearly aggregate the difference vector according to a topic weight vector  $\mathbf{w} \in \mathbf{R}^k$  and an intercept  $b$ , where  $\mathbf{w}$  and  $b$  are optimized through minimizing the Least Square Error between  $r_i^s$  and human annotations. The similarity measure  $r_i^s$  is formally defined as:

$$r_i^s = \mathbf{w}^T(\mathbf{topic}_i^f - \mathbf{topic}_i^t) + b.$$

### 3.3.3 Methods Performance Comparison

We evaluate the performance of each method against human annotations in a test data set. Recall that three annotators independently rate the level of congruence for each image and test pair, the average of which is used as the ground truth. For each pair, we apply our method and the two benchmark methods and obtain the correlation between the output and the ground truth. We observe that the cosine similarity method based on Benchmark Method 1 has the lowest correlation (0.35) among the three methods. In practice, this method tends to be the most commonly adopted because its building blocks, Google Vision API and LDA, are readily available and easy to implement. In Benchmark Method 2, we add the learning of topic weights in a supervised manner and increase the correlation to 0.56. The correlation between our proposed method and the human ground truth is 0.74. This confirms that our approach has a performance superior to current common methods. Note that our proposed method does not rely on large-scale human annotations, so it is scalable with respect to monetary and time costs. Table 3 presents a summary of the method comparison. More details about the method specification and implementation are presented in the Appendix.

---

Insert Table 3 about here.

---

## 4 Data and Model

### 4.1 Empirical Setting

Data for this research are provided by a Chinese leading platform company specializing in K-12 extracurricular online reading. The company offers a mobile app providing ebooks and audio-books for elementary and middle school students. Users pay a one-time subscription fee to consume the reading content. Once subscribed, users have unlimited access to all materials provided on the app. No other fees or in-app purchases are needed to consume the content.

In the market for young users, one may wonder who makes the product choice: the young users who consume the content or their parents who pay for the app. According to a customer survey conducted by the company, the young readers choose the content they read. The usage pattern from our data suggests this is true. We find that peak usage occurs around dinner and bedtime, when parents are busy with household chores such as cooking and cleaning (see Figure A4 in the Appendix). During the day, usage peaks around noon and 1 PM, when children are at school during lunch break. This provides suggestive evidence that it is the young users who make the product choice on the app.

## 4.2 Sample Description

Our sample includes consumption data from 15,966 unique users over a period of seven months from June to December, 2019. Table 4 presents the summary statistics. Across all users, the average age is 10.4 (SD = 2.3) years, with 49.8% being boys. Approximately 92.5% of the subjects are Android users (SD=0.26). At the beginning of the data collection period, the users had an average tenure of 3.74 months (SD = 1.88) on the platform.

---

Insert Table 4 about here.

---

A unique aspect of our data is that we can observe the complete time series of consumption for a user since s/he joins the app. Our observations are organized by “product session.” Each session corresponds to an incidence during which a user consumed the content of a product; i.e., a reading material. The materials on the app fall into two categories: audio-books (N=1,759) and ebooks (N=630). A product session begins when a user starts to browse a product and ends when the user starts to browse another product or becomes inactive for more than 5 minutes. If a user consumed the same audio-book or ebook multiple times, for example, reading different sections of the product or reading the product again, we analyze only the first time when the product was chosen and exclude future repeated consumption of the same product.

Table 4 presents the summary statistics for the sessions. There were 138,920 product sessions in our data, out of which roughly 53.4% (= 74,196/138,920) are audio-book sessions and the remaining 46.6% (= 64,724/138,920) are ebook sessions. Each session on average lasts 16.31 minutes, with audio-book sessions lasting longer than ebook sessions. On average, each user in our sample has 8.72 sessions (SD=12.04) and each person consumes audio-books slightly more than ebooks. It is

reasonable to assume that users’ behaviors may evolve as they become more familiar with the app and with the products. To capture this, we will allow users’ preferences of product choice to evolve over time, which we describe in the modeling section next.

One might wonder how the image-text congruence of an audio-book or ebook is related with its propensity of being chosen by consumers. We split products into nine congruence segments (from low to high, with the range of each being 0.05). For each congruence segment, we calculate the average consumption incidences across all the products within the segment, and plot the averages in Figure 3. An interesting U-shape is present in our data pattern: products with high or low image-text congruence on average had a higher propensity of being chosen, whereas those with a medium level of image-text congruence had a lower average consumption incidence. This model-free evidence has not yet adjusted for the many other control variables that could also influence people’s choice. Nevertheless, it provides motivational evidence for us to formally model a non-linear relationship between image-text congruence and consumer preference, which we present in detail next.

---

Insert Figure 3 about here.

---

### 4.3 Modeling Approach

We now describe our empirical approach to examine how image-text congruence affects users’ choice of products. It is assumed that user  $i$  derives utility  $U_{ijt}$  for product  $j$  at time  $t$ , which is modeled in the following specification:

$$\begin{aligned}
 U_{ijt} &= \beta_{ijt} \text{Congruence}_j + \gamma_{ijt} \text{Congruence}_j^2 \\
 &\quad + \eta_1 X_{it} + \eta_2 X_i + \eta_3 X_j + \eta_4 \text{Image}_j + \eta_5 \text{Text}_j + \alpha_i + \varepsilon_{ijt} \\
 \beta_{ijt} &= \beta_1 + \beta_2 X_{it} + \beta_3 X_i + \beta_4 X_j \\
 \gamma_{ijt} &= \gamma_1 + \gamma_2 X_{it} + \gamma_3 X_i + \gamma_4 X_j \\
 \alpha_i &= \alpha + \nu_i, \quad \nu_i \sim N(0, \sigma^2)
 \end{aligned} \tag{1}$$

where  $\text{Congruence}_j$ , the key variable of interest, is the congruence score between product  $j$ ’s profile image and its text description and  $\text{Congruence}_j^2$  is the quadratic term to capture a non-linear relationship between congruence and utility. Following the convention, we perform mean-centering on  $\text{Congruence}_j$  to reduce potential multicollinearity between the linear and quadratic term. The parameters  $\beta_{ijt}$  and  $\gamma_{ijt}$  capture the effects of image-text congruence on utility, which are allowed to



vary by session  $X_{it}$ , user characteristic  $X_i$ , and product characteristic  $X_j$ , respectively. Specifically,  $X_{it}$  is the product session for user  $i$  (1 is his/her first consumption session, 2 is the second, and so on...);  $X_i$  is the vector of user characteristics including age and gender; and  $X_j$  is the indicator for the product category ( $X_j = 0$  for ebooks and 1 for audio-books).

The parameter  $\alpha_i$  captures the random heterogeneity among users, which is assumed to be normal with mean  $\alpha$  and variance  $\sigma^2$ . The idiosyncratic errors,  $\varepsilon_{ijt}$ , are assumed to be independently and identically distributed (i.i.d) following the extreme type I distribution. Thus, user  $i$  at time  $t$  chooses whichever product  $j$  that maximizes the utility among all alternatives in the choice set; i.e.,  $j = \operatorname{argmax}_h U_{iht}$ , for  $h = 1, 2, \dots, I_t$ .

It is worth noting how user  $i$ 's choice set at time  $t$ ,  $I_t$ , is defined for the analysis. By definition,  $I_t$  contains all the products available on the app when user  $i$  makes a consumption decision at  $t$ . The choice set contains a large assortment of products and thus is computationally challenging to handle in the discrete choice model framework. To make the estimation feasible, we construct a small, but reasonable, choice set following the negative sampling method proposed by Goldberg and Levy (2014), which has become a popular method to create choice sets in the machine learning literature. For every product chosen in our data, the choice set contains 50 products not chosen by user  $i$  at time  $t$ .<sup>2</sup> The products in the choice set are not chosen at random; instead, they are the top 50 most consumed products on day  $t$ . Like most media apps, the focal app also lists popular products on its loading page. Thus, popular products are more likely to enter a user's consideration set than unpopular ones. Furthermore, using popular products would yield a more conservative estimate for the congruence effect, our main parameter of interest. Using a random sample to construct the choice set, however, may run the risk of overestimating the congruence effect.

Finally, in addition to the congruence between the product's profile image and the text description, users could also be affected by the media characteristics of the image and the description (denoted as vectors  $Image_j$  and  $Text_j$ , respectively). We control for these important variables and describe each one in details next. (See Table 6 for a summary of parameters and variables in Model 1.) Note that  $Image_j$  variables are also mean-centered to reduce potential multicollinearity.

---

<sup>2</sup>An analysis using sizes of 10 and 25 yielded qualitatively similar results.

### 4.3.1 Image Characteristics

We first construct visual variables to capture the aesthetic properties of a product’s profile image. Previous literature on image visual aesthetics has identified that the *color composition* of an image can affect people’s emotional responses. Valdez and Mehrabian (1994) found that *colorfulness*, including color saturation and lightness, drives pleasure and arousal. Marketing research has shown that consumers’ attitude towards advertisements can also be affected by visual aesthetic features such as the *color contrast* of an image (Finn 1988, Wedel and Pieters 2015). Deng et al. (2010) also found that while consumers in general prefer a small number of common colors, a contrasting color that highlights a single distinctive element of a product design may stand out and affect consumer preference.

**Colorfulness:** Following Li and Xie (2020), we process our images using Google Cloud Vision API, which extracts up to ten colors per image along with the corresponding pixel fraction (i.e., the fraction of area occupied by each color). From the extracted color elements, we generate two variables to measure the color composition of each image: *colorfulness* and *color contrast*. A picture’s colorfulness is calculated as one minus the sum of the pixel fraction across the five colors with the highest fraction. Thus, the variable colorfulness measures the variety of an image’s color composition, with a higher value corresponding to a higher degree of color variety. For example, if an image contains just two colors, its colorfulness is zero (i.e., not colorful). In contrast, if an image contains ten colors, each with 10% pixel fraction, the colorfulness equals  $1 - 0.1 \times 5 = 0.5$ , which is more colorful than the image with just two colors.<sup>3</sup> In our sample, the average colorfulness is 0.55 (SD = 0.17) for audio-books and 0.47 (SD = 0.11) for ebooks.

**Color Contrast:** We define an image’s color contrast as the degree to which the colors “stand out” from each other. To construct this metric, we again use the top five colors in each image<sup>4</sup> and compute the similarity between every pair of colors. The color similarity is defined as the cosine value of two vectors representing the color’s RGB code: a cosine value closer to 1 corresponds to two very similar colors and vice versa for a smaller cosine value. Thus, an image’s color contrast is

---

<sup>3</sup>Note that this “colorfulness” measure weights each color by its pixel fraction and thus is better than a simple count of distinct colors. Consider image A and B, both having ten colors. Imagine image A is dominated by two colors and the top five colors make up 98% of the pixel area, yielding a colorfulness value of 2%. Suppose image B has an even split among the ten colors, which yields a colorfulness value of 10%. Therefore, our definition of colorfulness can reflect the fact that image B is more colorful than image A.

<sup>4</sup>We did not use all ten colors identified by Google Cloud Vision API, because not all images have up to ten colors.

defined as one minus the average cosine similarity between all color pairs (i.e.,  $10(= 5 \times 4/2)$  pairs out of the top 5 colors). A higher color contrast value indicates high contrast and low similarity between the major colors of the image. The average color contrast is 0.11 (SD = 0.13) for audio-books and 0.06 (SD = 0.04) for ebooks.

**Image Objects:** In addition to the color composition, the *objects in an image*, such as the presence of a human face, may also affect consumer attitude (Li and Xie 2020). We again used Google Cloud Vision API to extract the objects contained in each image. The API compares an image’s pixel components to pre-tagged patterns and yields up to 10 object labels per image. Three independent human coders classified all the labels into object groups, depending on whether the labels refer to objects from the same category such as buildings, animals, and nature objects. The reliability among the coders was high (the Fleiss’ Kappa is 0.815,  $z = 52.1$ ,  $p < 0.001$ ). Any disagreements among the coders were resolved through discussion. For each image, we summarize the content using the number of labels and the percentage of labels belonging to each of the four largest categories: animal figures, human faces, nature objects, and emotions. See Table 5 for the summary statistics for all visual variables.

---

Insert Table 5 about here.

---

#### 4.3.2 Text Content of Product Description

We also control for the text content of a product’s description. On average, each product description contains 80.2 words with a standard deviation of 66.3 words. We adopt the topic extraction idea to summarize the description because the generated topics can capture the underlying association between words. To do so, we removed the stopwords and performed LDA model. The method takes the tokenized text segments (in our context, keywords) as the input and, for each word, estimates the distribution over a set of topics. LDA has the advantage of identifying common themes from a large number of unique keywords so that the content of the abstracts can be summarized by a modest number of topics. In our application, ten topics are identified across audio-books and ebooks. For the purpose of our analysis, we do not assign a meaning for each topic. Rather, the distribution over topics for each abstract directly enters our model as the control variables. To sum up, the control variables for our model include the image characteristics and text topics. In Table 6 we include the a full list of these variables with a short description of each.

---

Insert Table 6 about here.

---

## 5 Results and Discussion

In this section, we present our parameter estimates and discuss the potential mechanisms.

### 5.1 Parameter Estimates

We first describe the estimates for the main-effect model (Equation 1), which correspond to the average effect for a typical product. The results are listed as Model 1 in Table 7. The first thing to note is that image-text congruence plays an important role in consumer choice of audio-books and ebooks. Both the linear effect (1.717,  $p < 0.001$ ) and the quadratic effect (5.216,  $p < 0.001$ ) are estimated to be significant, indicating a non-linear relationship between image-text congruence and preference.<sup>5</sup> The positive quadratic effect suggests that, everything else being equal, user preference towards a product is higher when there is either a low degree or high degree of congruence between the product’s profile picture and its text description. This non-linear effect further varies by session, product type (audio-books versus ebooks), and user characteristics such as age and gender (see Model 2 in Table 7). In Section 5, we provide more details on this effect and discuss potential mechanisms.

Next, the image’s visual features are found to affect consumer choice. More colorful cover images are more likely to attract consumers (0.262,  $p < 0.001$ ), and those with higher color contrast (i.e., with colors that are less congruent with each other) also have a positive effect on choice (0.065,  $p < 0.05$ ). In terms of image content, our results show that the number of detected objects is positively associated with choice (0.056,  $p < 0.001$ ): on average, presenting more objects in the cover image helps choice. Among the types of objects, consumers respond more positively to animals (1.622,  $p < 0.001$ ), human faces (0.847,  $p < 0.001$ ), and emotions (0.809,  $p < 0.001$ ), but negatively to nature objects (-0.195,  $p < 0.001$ ), which perhaps reflects consumers’ reading preferences across genres. Lastly, products’ text descriptions also matter for consumer choice. The topics identified from LDA analysis are all found to be positively related with consumer choice (relative to the reference topic), although the magnitude varies.

---

Insert Table 7 about here.

---

<sup>5</sup>Note that the congruence variable is mean-centered, with the value ranging between -0.40 and 0.25 in the analysis.

## 5.2 Results Discussion and Mechanism Exploration

We now focus on our key construct of interest: the image-text congruence. As summarized in the above section, consumers are more likely to choose a product when the image-text congruence is either high or low. In contrast, everything else being equal, consumers prefer the product the least when its image-text congruence is at a medium level, resulting in an conspicuous “dull zone.” This U-shape effect seems to suggest that the congruence between the two ubiquitous stimuli influences consumers in a rather complex way. However, before we explore the potential mechanisms for congruence, one might wonder to what extent consumers pay attention to these stimuli when they choose entertainment materials (i.e., audio-books and ebooks in our context). For example, consumers may pay little attention to either the image or text and hence the image-text congruence would not matter. This conjecture is ruled out by the aforementioned results of image characteristics and textual topics. The fact that image composition (colorfulness and color contrast), image content, and text content have significant effects indicates that consumers pay attention to these stimuli. Thus assured, we turn to information processing theories to understand the effect of image-text congruence.

### 5.2.1 Decomposing image-text congruence

Understanding why and how image-text congruence affects consumers is of critical importance for researchers and practitioners designing and optimizing product communications. We turn to the literature in consumer behavior and cognitive psychology to explore the potential underlying mechanisms. Research in information processing has examined multi-media advertising (involving print and video) and proposed two constructs that may contribute to the congruence between different media formats: *relevancy* and *expectancy* (Heckler and Childers 1992, Lee and Mason 1999). *Relevancy* is defined as the material pertaining directly to the meaning of the theme. When two sources have relevant content, the information contained in one stimulus contributes to (rather than distracts from) the theme or message being communicated in the other stimulus. When two stimuli are highly relevant, they evoke information processing fluency (Hastie 1980, 1981, Srull 1981, Srull et al. 1985), leading to increased consumer preference.

*Expectancy* refers to the degree to which information contained in different stimuli falls into some predetermined pattern evoked by the theme (Goodman 1980, Heckler and Childers 1992). When two information sources are high in *expectancy*, the content embedded in one would appear as expected

and unsurprising to consumers compared with the content contained in the other. Research in cognitive psychology and consumer behavior has found that compared to expected information, *surprise* can raise attention and thus lead to more elaborate information processing and encoding, which in turn can boost recall and become an underlying driver for preference (Heckler and Childers 1992, Lee and Mason 1999). This dichotomy framework unveils two underlying psychological drivers of consumer preference embedded in the concept of congruence: information processing fluency through *relevancy* and information elaboration through *surprise* (i.e., low expectancy).<sup>6</sup>

Following this literature, we focus on *relevancy* and *expectancy* (or reversely, *surprise*) as our key constructs to disentangle how these two components affect congruence between product pictures and textual descriptions and consequently influence consumer choice. In our context, *relevancy* measures how much the image contributes to the theme communicated in the product’s text summary and *expectancy* reflects the lack of “surprise” when consumers jointly process the information contained in the cover image and text. Imagine that a consumer sees a cover image and forms an expectation on what the book is about. If the book summary conforms to the expectation, the image-text pair would be high in *expectancy*. The same logic applies if the consumer processes the text stimuli first and bases the expectation on text. With these potential underlying constructs defined, we proceed to investigate how they drive congruence and affect consumer choice in our setting. Because expectancy affects consumer choice indirectly through the surprise effect, we focus our analysis and discussion on *surprise* hereafter to avoid repetition.

Figure 4 summarizes the proposed framework.

---

Insert Figure 4 about here.

---

### 5.2.2 Two-driver Hypothesis: Empirical Evidence from Browsing Data

So far, we hypothesize that *relevancy* and *surprise* are the two drivers of consumer preference and the underlying mechanism for our findings. If our postulation holds, one would expect that consumers would spend more time deciding when they choose products of low image-text congruence than those of high congruence. After all, it takes less effort to process relevant information to construct a common theme (Hastie 1980, 1981, Srull 1981, Srull et al. 1985).

---

<sup>6</sup>Note that although these two constructs could be related to each other, they each contribute orthogonal information. The phrase “business causal” and an image of a pink jacket would likely be high in *relevancy* but low in *expectancy*. In contrast, the word “meatball” and an image of spaghetti could be high in *expectancy* but may be low in *relevancy*.

To examine this hypothesis, we collect the time that consumers spent on browsing before they chose each product. Figure 5 depicts the average browsing time leading to a product choice by congruence level. For the products with low image-text congruence, it takes the consumers twice as long to make that choice. One-way ANOVA with Tukey adjusted post-hoc tests indicate that the difference is significant between the low congruence group and the middle and high congruence groups. The browsing time, however, is not statistically different between mid and high congruence levels. These results provide suggestive evidence that, at low image-text congruence, *surprise* seems to drive consumer preference and offsets the negative impact of low *relevancy* because consumers indeed elaborate more in this case.

---

Insert Figure 5 about here.

---

### 5.2.3 Online Study Design

Although our two-driver hypothesis is supported indirectly by the observed pattern of browsing behaviors, we would like to pin down the mechanism in a controlled study. To do so, we conduct an online study to measure *relevancy* and *surprise* between the image and text pairs. A total of 144 college students are recruited to a website specifically created for this study. Each student is assigned 100 image and text pairs randomly selected from our data sample. For each pair, the participant is asked the following questions: 1) Based on the text summary and the cover image of a book shown, how much do you think the book cover is relevant with the text (1 as “completely irrelevant”, 7 as “very much relevant”); 2) Based on the text summary and the cover image of a book shown, how much do you think the book cover is as expected (1 as “completely unexpected”, 7 as “very much expected”); and 3) Based on the text summary and the cover image of a book shown, how much are you curious to read the book? (1 as “not at all curious”, 7 as “very much curious”). The first two questions measure *relevancy* and *surprise*,<sup>7</sup> respectively, and the last question measures consumers’ consumption intention. A screenshot of the online survey interface is included in the Appendix.

On average, each image and text pair is rated by five respondents. Their averages make up the final score of *relevancy* and *surprise* scores for each product. For *relevancy*, the mean across all products is 4.48, with a standard deviation of 0.88. The average for *surprise* is 3.76, with a standard deviation of 0.85. Table 8 presents the summary statistics of the survey responses.

---

<sup>7</sup>As described in the previous section, *surprise* is a more direct driver of behavior and more straightforward in our mechanism analysis. Therefore, we reverse-coded the *expectancy* variable to construct *surprise*.

---

Insert Table 8 about here.

---

#### 5.2.4 Mediation Analysis

To examine how *relevancy* and *surprise* are related to the congruence metric, we categorize the products into three groups: the “Low” group is the products whose congruence measure is in the bottom 35% , the “High” group is the top 35%, and the remainder is the “Medium Congruence” group. Figure 6 depicts the average *relevancy* and *surprise* scores by congruence group. As expected, *relevancy* and *surprise* exhibit different patterns with congruence: whereas the average *relevancy* score increases with congruence, the average *surprise* level decreases as congruence increases. In other words, products with a high level of congruence tend to be high in *relevancy*, but low in *surprise*. One-way ANOVA with Tukey post-hoc tests indicate that the difference is statistically significant for every pairwise comparison.

---

Insert Figure 6 about here.

---

The descriptive pattern in Figure 6 establishes that *relevancy* and *surprise* vary significantly among the low, middle, and high congruence groups. However, it does not answer the question of whether *relevancy* and *surprise* are the underlying mechanisms that drive the impact of congruence on consumer choice. To provide direct evidence, we conduct a mediation effect analysis. The logic goes as follows. Our main results in Table 7 and Figure ?? show the U-shape relationship between congruence and consumer utility. If *relevancy* and *surprise* are indeed the underlying drivers for consumer preference, one would expect the effect of congruence to diminish or completely disappear after *relevancy* and *surprise* are controlled for in the model.

We perform the mediation test on the survey data following the Baron-Kenny approach (Baron and Kenny 1986). In step 1, we regress consumer intention on the product congruence measure to obtain the total effect. Results confirm the U-shape curve: the coefficient for the linear term is estimated to be 1.311 ( $p < 0.001$ ) and the coefficient for the quadratic term is also positive and significant at 5.050 ( $p < 0.001$ ). Model 1 in Table 9 summarizes the results.

In step 2, we add *relevancy* and *surprise* into the model and obtain the remaining direct effect of congruence on preference. To do so, we calculate, for each product, the average score of *relevancy*



and *surprise*, which reflects a congruence measure based on the survey construct. Results in Table 9 Model 2 indicate that the composite survey construct has a significant U-shaped relationship with preference (i.e., coefficient for Survey Construct Square is 0.050,  $p < 0.001$ ). After controlling for the survey construct, however, the direct impact of congruence is no longer significant. In other words, the survey constructs seem to completely mediate the effect of congruence.<sup>8</sup>

---

Insert Table 9 about here.

---

### 5.2.5 Mechanism Discussion

Our results so far have suggest the following findings: (1) consumers respond more positively when image and text congruence is high or low; and (2) the impact of congruence seems to be mediated by the level of *relevancy* and *surprise* between the two media. These findings therefore imply that an optimal level of congruence depends on the combination of *relevancy* and *surprise*. When the product image and text descriptions are high in *relevancy* and low in *surprise* (i.e., high in congruence), the high consistence between the two stimuli could evoke information processing fluency (Hastie 1980, 1981, Srull 1981, Srull et al. 1985), leading to increased consumer preference.

More interestingly, the positive impact of low congruence is less intuitive and worth discussing. When image and text are low in *relevancy* but high in *surprise*, consumers will perceive the two stimuli as low in congruence. Although low *relevancy* is negatively associated with utility, the *surprise* element could have a positive effect: high *surprise* can raise attention and lead to more elaborated information processing and encoding, which in turn can boost recall and become an underlying driver for preference (Heckler and Childers 1992, Lee and Mason 1999). These two effects work through different routes, leaving the middle congruence level as a “dull zone” for product preference.

## 6 Conclusion

Firms have been using multiple media stimuli in advertisements and product displays for decades. On online platforms now, it is increasingly common for a product’s text description to be associated with visual images and even audio and video presentations. Naturally, the congruence between

---

<sup>8</sup>To demonstrate the robustness of this result, we also compute an alternative version of the survey construct by allowing *relevancy* and *surprise* to have different weights contributing to congruence. We find consistent results as in Table 9.

multiple media stimuli could play a significant role in affecting consumers’ attitudes and choices. However, due to the difficulty of analyzing unstructured data (e.g., images and text), little empirical research has been conducted to measure the congruence between image and text, let alone to examine how image-text congruence affects consumer information processing. We fill this important gap.

In this study, we propose an innovative method to measure the semantic congruence between an image-text pair. Our method utilizes state-of-the-art image and text embedding techniques to extract the semantic meaning from each medium. The embeddings then enter a two-branch supervised deep learning model, where the parameters are optimized to predict the level of image-text congruence (against human annotation). The innovation of our method is that the image and text embeddings are jointly modeled, yielding a performance superior to that from existing popular methods that model images and text separately. Furthermore, our method outputs a continuous measure of congruence, which makes it possible to capture the variation in image-text congruence in a reliable and scalable manner.

We apply the congruence measure to the granular data from a digital platform specializing in online reading for young users. Using the consumption records of ebooks and audio-books for more than 11,000 users, we build a random-coefficient logistic model to examine the effects of image-text congruence on consumer product choice, after controlling for other factors such as image features, text topics, time, and consumer heterogeneity. Our results reveal an interesting U-shape relationship between image-text congruence and consumer preference: consumers are more likely to choose a product when the congruence between its cover image and text description is either high or low. However, the middle level of image-text congruence is associated with the lowest consumer preference, constituting a “dull zone”.

To explore the mechanism of the congruence, we propose two underlying drivers through which the congruence between multi-media stimuli exert influence on consumer information processing: 1) information fluency when there is high relevancy between the content of the image and that of text, and 2) surprise-induced information elaboration when the content of one media type is unexpected based on the content of the other. We conduct an online study and use survey questions to measure the constructs of *relevancy* and *surprise* for a random sample of products. Our results confirm that products of high image-text congruence are high in *relevancy* and that those of low image-text congruence are high in *surprise*. Furthermore, *relevancy* and *surprise* are both positively associated

with consumer preference. When the influence of *relevancy* and *surprise* is controlled for, image-text congruence is no longer significant for consumer preference. These results suggest that the influence of image-text congruence is completely mediated through *relevancy* and *surprise*.

## 6.1 Theoretical Implications

Our research contributes to the literature of multi-media stimuli in consumer information processing. While conventional wisdom suggests that the “fit” between image and text matters in practice, the question of how to measure image-text congruence and understand the subsequent effect on product choice has largely been absent in the literature. Through our mechanism analysis, we pin down the drivers underlying the heuristic concept of “congruence.” Our findings indicate that “congruence” between image and text is more complicated than “relevancy” or “fit,” although the latter is a key component of the former. When consumers process information from multiple sources, they form an expectation based on the initial source and cross-check the information when encountering the second source. If the new information is somewhat unexpected, consumers spend more time processing the “surprise,” leading to a more elaborated processing state. Our results confirm that this elaborated information processing can also lead to more positive outcomes. To the best of our knowledge, this research is among the first attempts to identify a U-shape relationship between media congruence and consumer preference. While consumer behavior literature has theorized a positive surprise effect and found some initial evidence in labs, we provide much-needed empirical evidence and validate this effect using actual consumption data at a much larger scale.

## 6.2 Managerial Implications

Our study offers several important managerial implications for both product managers and content creators. First, for digital marketers, our findings can help identify the optimal level of image-text congruence in product design and product communications. We show that it is important to recognize there are two dimensions of congruence: *relevancy* and *surprise*. Managers can consider pairing the images and text when their content has high *relevancy* and low *surprise* or pair the content with moderate *relevancy* and high *surprise*. This finding offers concrete guidelines on how to design the underlying content elements to achieve an optimal congruence level.

Second, the fact that all effects fade over time sheds light on consumer segmentation strategies. In our setting, the congruence between images and texts are most effective for new users. As they become more familiar with the products on the platform, the cover design seems to lose its

influence. We speculate that experienced users are more likely to infer content from the book’s title and summary; as a result, the congruence between the cover image and the text summary no longer matters as much. Thus, managers should pay special attention to image and text congruence when they recommend products to newer customers.

Third, our study can also help product managers understand the effect of image features such as colorfulness and color-contrast. For example, we find that products benefit from more colorful designs as long as the color mix has a reasonable contrast level. Last, but not least, our innovative measure of image-text congruence offers a useful metric to study misinformation on social media. For example, some online content creators use an attractive image to lure viewers, while the content of the image may mismatch or have low relevance with the true story in text, thus creating the “click-bait” phenomenon. Our technical framework can be the first step in tackling this issue. Platforms can adopt this method to better detect and identify the suspected “click-baits” in a more efficient way, which helps increase customer satisfaction and users engagement.

### **6.3 Limitations and Future Research**

Our study can benefit from several future research directions. First, we only study ebooks and audio-books from one market: online reading for young users. Although we believe our congruence metric and its mechanism can be generalized to other settings, our findings can benefit from validation in other settings, especially retail platforms (other than books), social media, and advertising. Second, since the images and text in our sample are firm-generated, we do not include extreme cases where the information delivered from the two sources are completely irrelevant. However, this extreme case might exist in user-generated content. It would be interesting to apply our method to user-generated content of social media posts and product reviews. We expect the effect of image-text congruence to vary by the type of user decision. For example, the surprise effect might be less pronounced for user reviews than for media consumption. Nevertheless, this study offers a metric to study the effect in various future settings. Third, we only focus on images and text descriptions, while more multi-media stimuli such as audio and videos can be analyzed in the same spirit. Although the means of extracting semantic meanings from different media formats may vary, we believe our two-branch modeling framework can be generalized and extended to other media formats. In sum, this research lays a foundation for future research in this area, which combined will contribute to the much-needed knowledge on consumer processing of multiple media.

## References

- Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51(6):1173.
- Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. *Proceedings of ICML workshop on unsupervised and transfer learning*, 17–36 (JMLR Workshop and Conference Proceedings).
- Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the tribes: Using text for marketing insight. *Journal of Marketing* 84(1):1–25.
- Berger J, Milkman KL (2012) What makes online content viral? *Journal of marketing research* 49(2):192–205.
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.
- Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2019) Uniter: Learning universal image-text representations .
- Cui Y, Che W, Liu T, Qin B, Wang S, Hu G (2020) Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922* .
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee).
- Deng X, Hui SK, Hutchinson JW (2010) Consumer preferences for color combinations: An empirical analysis of similarity-based color relationships. *Journal of Consumer Psychology* 20(4):476–484.
- Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .
- Dew R, Ansari A (2018) Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science* 37(2):216–235.
- Finn A (1988) Print ad recognition readership scores: An information processing perspective. *Journal of Marketing Research* 25(2):168–177.
- Ge X, Chen F, Jose JM, Ji Z, Wu Z, Liu X (2021) Structured multi-modal feature embedding and alignment for image-sentence retrieval. *arXiv preprint arXiv:2108.02417* .
- Goldberg Y, Levy O (2014) word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* .

- Goodman GS (1980) Picture memory: How the action schema affects retention. *Cognitive Psychology* 12(4):473–495.
- Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16(12):2639–2664.
- Hastie R (1980) Memory for behavioral information that confirms or contradicts a personality impression. *Person memory (PLE: Memory): The Cognitive basis of social perception* 155–178.
- Hastie R (1981) Schematic principles in human memory. *Social cognition: The Ontario Symposium*, volume 1, 39–88.
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Heckler SE, Childers TL (1992) The role of expectancy and relevancy in memory for verbal and visual information: what is incongruity? *Journal of consumer research* 18(4):475–492.
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25:1097–1105.
- Lee YH, Mason C (1999) Responses to information incongruity in advertising: The role of expectancy, relevancy, and humor. *Journal of consumer research* 26(2):156–169.
- Li Y, Xie Y (2020) Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research* 57(1):1–19.
- Liu X, Singh PV, Srinivasan K (2016) A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science* 35(3):363–388.
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .
- Pieters R, Wedel M (2004) Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of marketing* 68(2):36–50.
- Ramos J, et al. (2003) Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, volume 242, 29–48 (Citeseer).
- Shedler J, Manis M (1986) Can the availability heuristic explain vividness effects? *Journal of personality and social psychology* 51(1):26.
- Srull TK (1981) Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human learning and memory* 7(6):440.

- Srull TK, Lichtenstein M, Rothbart M (1985) Associative storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11(2):316.
- Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. *International conference on artificial neural networks*, 270–279 (Springer).
- Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Science* 38(1):1–20.
- Valdez P, Mehrabian A (1994) Effects of color on emotions. *Journal of experimental psychology: General* 123(4):394.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in neural information processing systems*, 5998–6008.
- Wang L, Li Y, Huang J, Lazebnik S (2018) Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2):394–407.
- Wedel M, Pieters R (2015) The buffer effect: The role of color when advertising exposures are brief and blurred. *Marketing Science* 34(1):134–143.
- Wei X, Zhang T, Li Y, Zhang Y, Wu F (2020) Multi-modality cross attention network for image and sentence matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.
- Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. (2016) Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .
- Xu X, Wang T, Yang Y, Zuo L, Shen F, Shen HT (2020) Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems* 31(12):5412–5425.
- Zagoruyko S, Komodakis N (2016) Wide residual networks. *arXiv preprint arXiv:1605.07146* .
- Zhang M, Luo L (2018) Can user-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Evidence from Yelp (March 1, 2018)* .
- Zhang S, Lee D, Singh PV, Srinivasan K (2021a) What makes a good image? airbnb demand analytics leveraging interpretable image features. *Forthcoming at Management Science* .
- Zhang S, Mehta N, Singh PV, Srinivasan K (2021b) Frontiers: Can an artificial intelligence algorithm mitigate racial economic inequality? an analysis in the context of airbnb. *Marketing Science* .

# Tables and Figures

Figure 1: Illustration of two-branch neural networks for image-text congruence

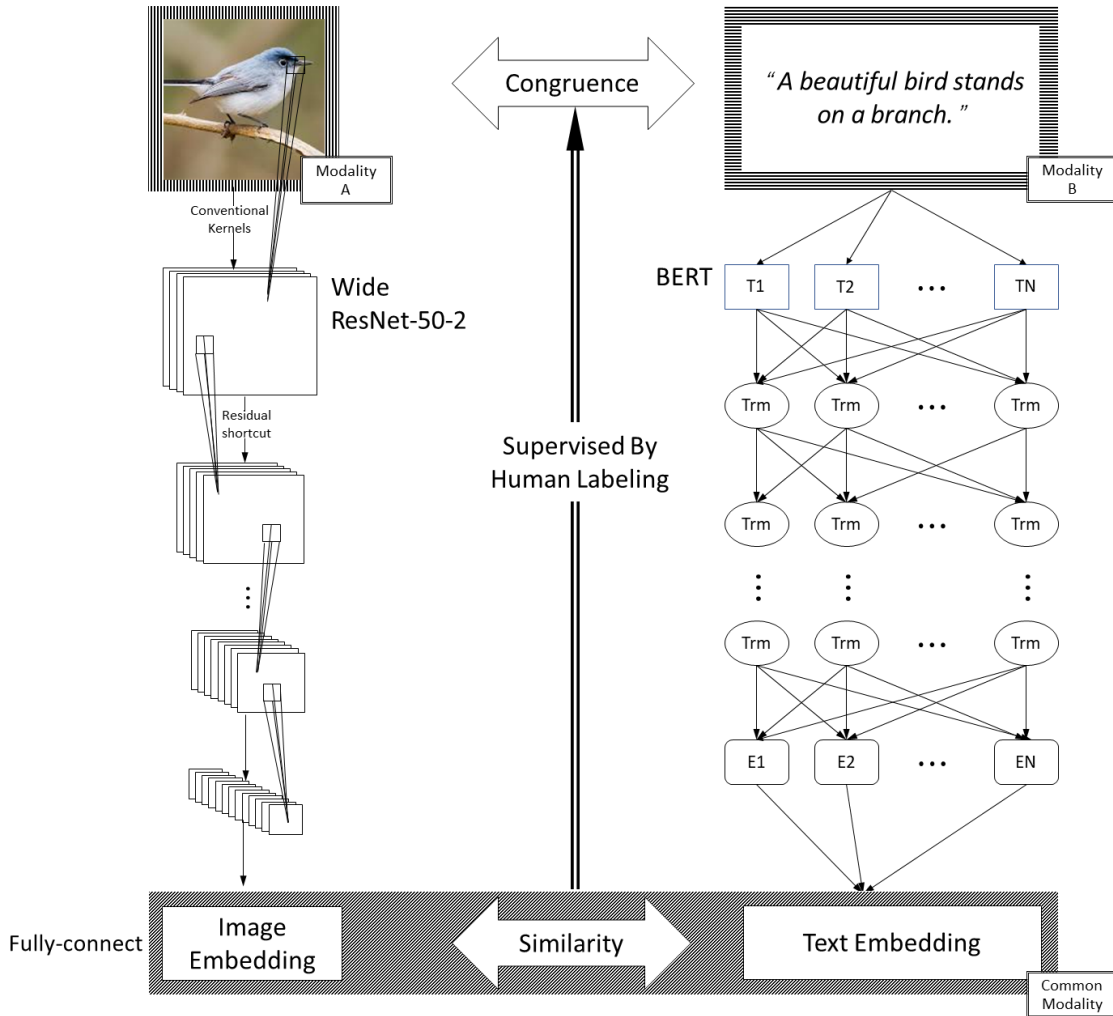




Figure 2: Mean Square Error (MSE) Performance for Training and Test Sets

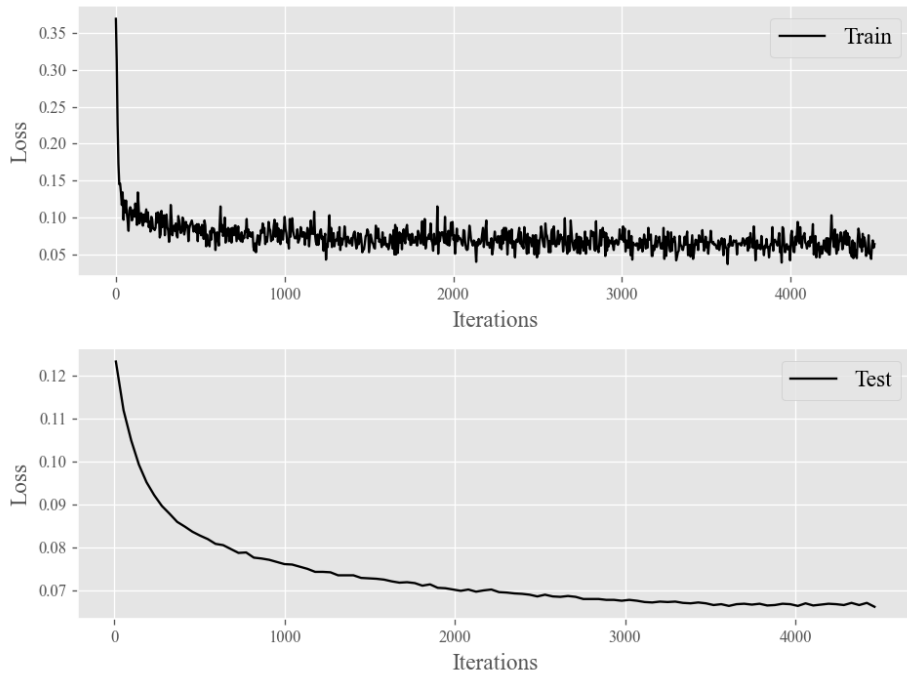
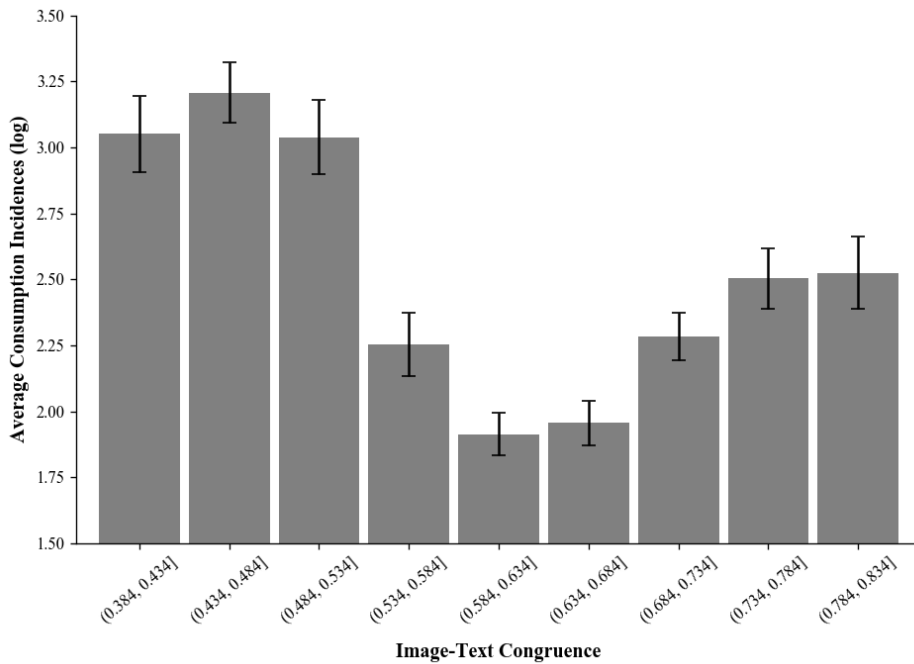


Figure 3: Model Free Evidence: The Image-Text Congruence vs. the Average Consumption Incidences



Note: The bars represent the average consumption incidences.

The top and bottom error bars correspond to plus and minus one standard error, respectively.

Figure 4: Mechanism Framework for Image-Text Congruence

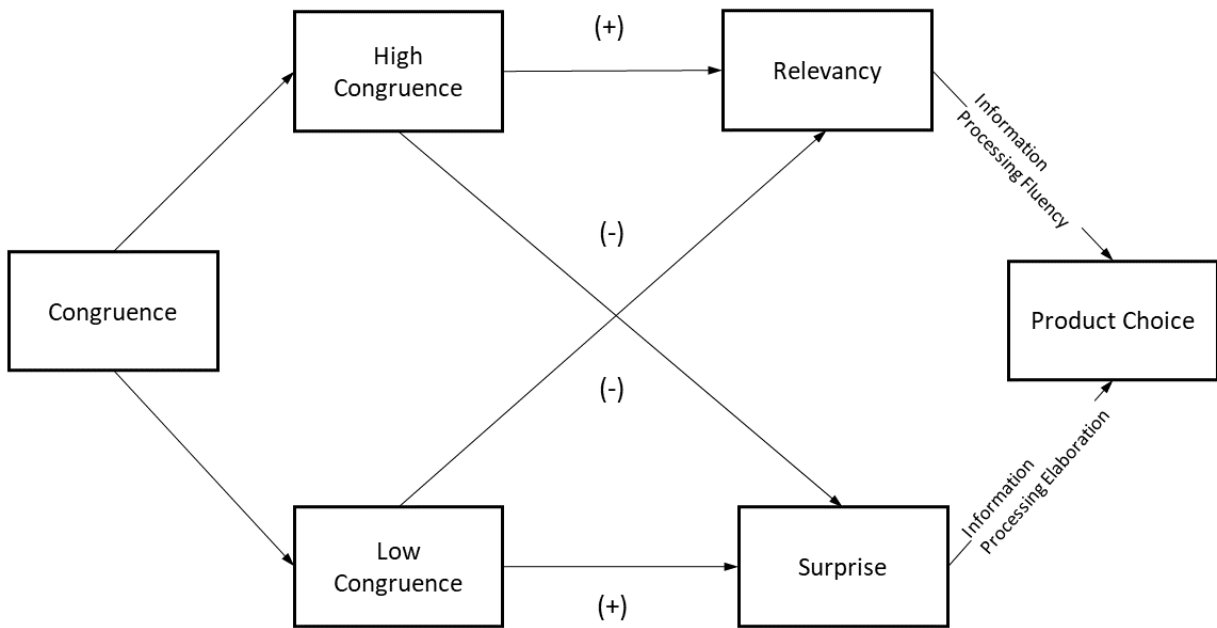


Figure 5: Browsing Time by Level of Image-Text Congruence

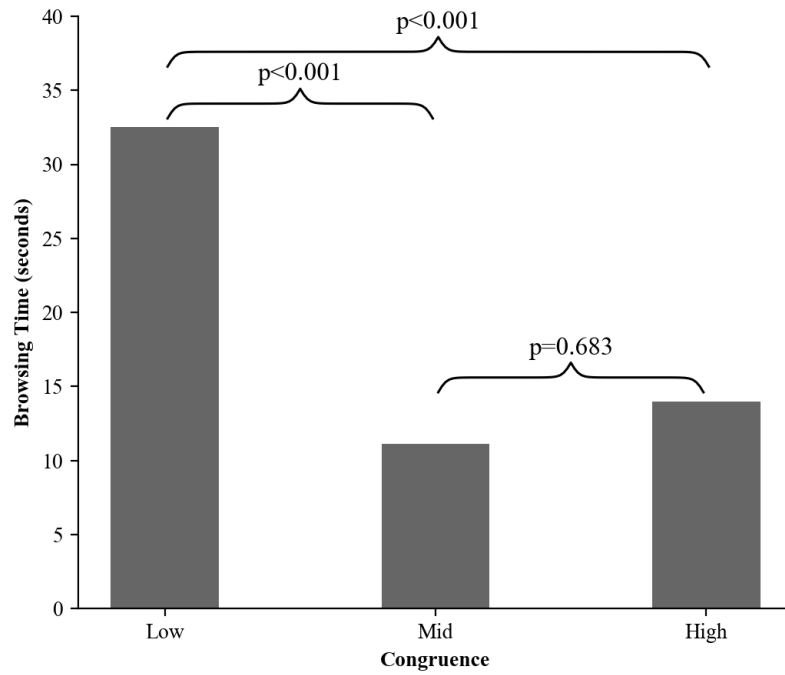
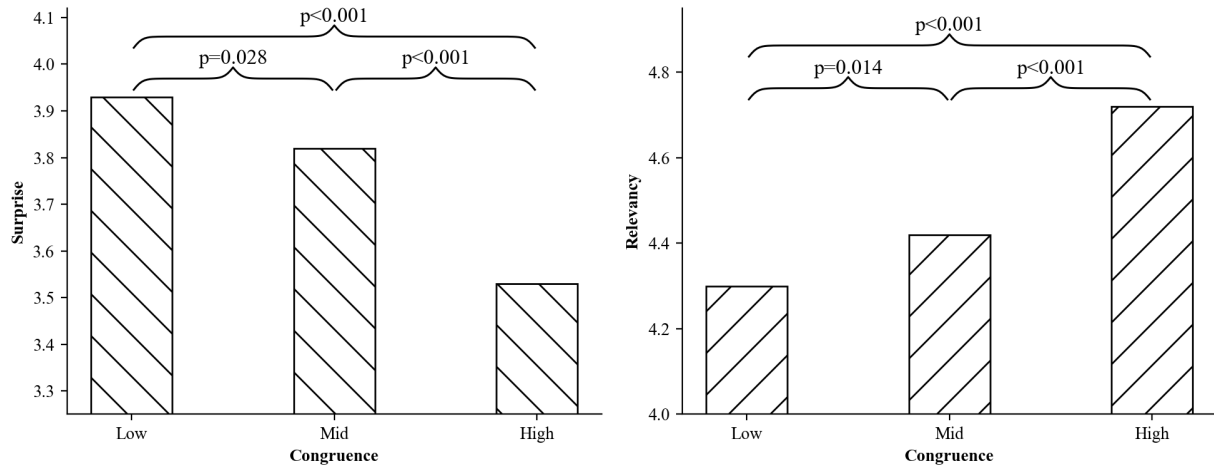


Figure 6: Survey Constructs by Level of Image-Text Congruence



Note: The left and right panels depict the mean *Surprise* value and *Relevancy* value by level of image-text congruence, respectively.

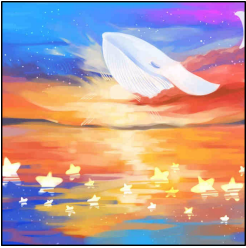


The p-values are from Tukey post-hoc comparisons following one-way ANOVA.

Table 1: Pearson Correlations between Annotators and Two-Branch Model

	Annotator1	Annotator2	Annotator3	Annotators Avg	Two-Branch Model
Annotator1	1	0.899	0.701	0.944	0.729
Annotator2	0.899	1	0.742	0.956	0.683
Annotator3	0.701	0.742	1	0.865	0.618
Annotators Avg	0.944	0.956	0.865	1	0.741
Two-Branch Model	0.729	0.683	0.618	0.741	1

Note: The unit of observation is an image-text pair.

Table 2: Congruence Measure for Three Example Image-Text Pairs

Image	Text	Congruence Measure
	<p><b>To Diuidu Who Refuses to Nap:</b>            Let's try not to disturb others' napping.            You know, they are very tired and need sleep badly.            How they wish to have matchsticks to keep their eyes open!</p>	0.140
	<p><b>One Hundred Thousand Questions Series: Origin of Life</b>            Where does life come from?            Besides animals and plants, what other lives are out there?            Why can bacteria keep self-reproducing?            This series has all the answers to what you are curious about.</p>	0.523
	<p><b>School of Elephants: A Trip to the Zombie Land</b>            Hello dear friends, the four troublemakers from the School of Elephants are back! What adventure do they have to share this time? Let's start reading!</p>	0.736

Note: The congruence measure was predicted by our two-branch DNN model.

Table 3: Method Comparison

	Ground Truth	Benchmark Method 1	Benchmark Method 2	Current Approach
	Human annotation	Unsupervised image-to-text method	Supervised image-to-text method	Supervised two-branch method
Method Outline	Human coders evaluate the congruence between image and text	Image and text are converted into vectors independently using an unsupervised approach.	Image and text vectors are assigned different weights to the distance calculation, which are learned through a supervised model.	A supervised model is fit by jointly finetuning the parameters of a two-branch pre-trained deep learning model.
Performance	1	0.35	0.56	0.74
Cost	Labor intensive; requires crowdsourcing	No manual labor needed; monetary costs on using Google Vision API; fast model inference and results delivery (less than 1 minute with 1,000 pairs)	No manual labor needed; monetary costs on using Google Vision API; fast model inference and results delivery (less than 1 minute with 1,000 pairs)	Limited manual labor; fast model inference and result delivery (less than 2 minutes with 1,000 pairs)
Scalability	Increasing marginal cost due to human fatigue; inflexible to apply to different tasks	Decreasing marginal cost with data scale; but heavily depend on the functions and scalability of Google Vision API; inflexible to apply to different tasks	Decreasing marginal cost with data scale; but heavily depend on the functions and scalability of Google Vision API; inflexible to apply to different tasks	Decreasing marginal cost; flexible to apply to different tasks with relatively low training cost (such as sentiment analysis of photos)
Stability	Results may vary by the quality of annotators and their working status	Results can be replicated as model is deterministic	Results can be replicated as model is deterministic	Results can be replicated as model is deterministic

Note: For supervised approach, the training/test split is 80%/20%.

Table 4: Descriptive Statistics of Users and Sessions

	N	Mean	SD	Min	Max
By User					
Age	15,966	10.42	2.26	5.42	16
Gender (1=Male, 0=Otherwise)	15,966	0.50	0.50	0	1
Operating System (1=Android, 0=iOS)	15,966	0.93	0.26	0	1
Tenure (months)	15,966	3.74	1.88	0	6.56
Number of Sessions	15,966	8.72	12.04	1	233
% Audiobooks	15,966	0.58	0.36	0	1
% eBooks	15,966	0.42	0.36	0	1
By Session: Duration (min)					
All Sessions	138,920	16.31	29.18	0.03	325.78
Audio-book Sessions	74,196	19.33	31.55	0.03	325.78
Ebook Sessions	64,724	12.85	25.77	0.03	325.52

Note: The users are the young readers using the App. The sessions refer to incidences of product consumption.

Table 5: Descriptive Statistics of Variables

	Audio-books N=1,759				Ebooks N=630			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Image-Text Congruence	0.71	0.07	0.34	0.88	0.53	0.12	0.23	0.87
Colorfulness	0.55	0.17	0	0.89	0.47	0.11	0.01	0.85
ColorContrast	0.11	0.13	0	0.91	0.06	0.04	0	0.34
# of Labels	6.56	2.67	1	10	8.02	1.28	1	10
% Animals	0.01	0.04	0	0.90	0.03	0.04	0	0.50
% Human	0.10	0.09	0	0.50	0.07	0.06	0	0.50
% Nature	0.02	0.09	0	0.90	0.08	0.08	0	1
% Emotion	0.02	0.08	0	0.50	0.02	0.02	0	0.40
Topic1	0.12	0.29	0	1	0.08	0.20	0	0.99
Topic2	0.12	0.30	0	1	0.07	0.20	0	0.99
Topic3	0.07	0.22	0	1	0.04	0.16	0	1
Topic4	0.21	0.36	0	1	0.17	0.32	0	0.99
Topic5	0.07	0.22	0	1	0.08	0.23	0	1
Topic6	0.04	0.15	0	1	0.05	0.18	0	0.99
Topic7	0.15	0.27	0	1	0.09	0.20	0	0.99
Topic8	0.09	0.26	0	1	0.11	0.26	0	1
Topic9	0.08	0.20	0	1	0.26	0.39	0	1
Topic10	0.04	0.16	0	1	0.06	0.19	0	0.99

Note: The unit of observation is a product (i.e., audio-books and ebooks).

Table 6: Summary of Variables and Parameters in the Main Model

	Notation	Definition and Measure
Variable of Interest		
	$Congruence_j$	Image-Text Congruence, continuous score (See Section 3)
Control Variable		
	$X_{it}$	Product Session per individual: 1, 2, ...
	$X_i$	Age: continuous scale Gender: 1 if male, 0 if otherwise
	$X_j$	Product Category: 1 if audio-books, 0 if ebooks
	$Image_j$	Image Aesthetics: Colorfulness, continuous scale Image Aesthetics: Color Contrast, continuous scale Objects in Image: % of nature objects, human characters, etc.
	$Text_j$	Topics in the text content
Parameters		
	$\beta_{ijt}$	Effects of image-text congruence (linear)
	$\beta_2$	Time-varying deviation by session
	$\beta_3$	Deviation by individual characteristics, time-invariant
	$\beta_4$	Deviation by product category, time-invariant
	$\gamma_{ijt}$	Effects of image-text congruence (quadratic)
	$\gamma_2$	Time-varying deviation by session
	$\gamma_3$	Deviation by individual characteristics, time-invariant
	$\gamma_4$	Deviation by product category, time-invariant
	$\alpha_i$	Random heterogeneity among users



Table 7: Parameter Estimates from Main Models

	Model 1		Model 2	
	EST	SE	EST	SE
Congruence Linear	1.717***	0.029	3.845***	0.115
Congruence Quadratic	5.216***	0.197	5.459***	0.908
Session	0.0003*	0.0001	0.001***	0.0002
Session×Congruence Linear			-0.029***	0.001
Session×Congruence Quadratic			-0.048***	0.009
AudioBook	-0.319***	0.008	-0.159***	0.009
AudioBook×Congruence Linear			-4.316***	0.070
AudioBook×Congruence Quadratic			1.129*	0.526
Age	-0.001	0.001	-0.014***	0.002
Age×Congruence Linear			-0.066***	0.010
Age×Congruence Quadratic			0.656***	0.078
Gender (Male=1)	-0.001	0.006	0.024**	0.008
Male×Congruence Linear			0.253***	0.042
Male×Congruence Quadratic			-2.415***	0.342
Colorfulness	0.262***	0.021	0.203***	0.021
ColorContrast	0.065*	0.029	0.061***	0.030
Number of Labels	0.056***	0.001	0.036***	0.002
% Animal	1.622***	0.060	1.741***	0.059
% Human	0.847***	0.041	0.667***	0.041
% Nature	-0.195***	0.040	-0.065+	0.040
% Emotion	0.809***	0.036	0.981***	0.036
Topic 1	0.344***	0.020	0.329***	0.020
Topic 2	1.015***	0.020	0.903***	0.020
Topic 3	0.587***	0.021	0.380***	0.022
Topic 4	0.085***	0.019	0.048*	0.019
Topic 5	0.245***	0.022	0.107***	0.022
Topic 6	0.323***	0.025	0.242***	0.025
Topic 7	0.175***	0.021	0.125***	0.021
Topic 8	0.231***	0.020	0.172***	0.020
Topic 9	0.263***	0.019	0.002	0.020
Intercept	-4.141***	0.023	-3.940***	0.027
Log Likelihood	-659,226.9		-657,041.4	
AIC	1,318,497.8		1,314,142.8	
BIC	1,318,604.0		1,314,287.7	

Note: The unit of observation is a choice incident. The number of observations is 6,736,488, from 15,966 users. Variable congruence is mean-centered, with the value ranging between -0.40 and 0.25 used in the analysis.

\*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ , + $p < 0.10$

Table 8: Descriptive Statistics of Variables

		N	Mean	SD	Min	Median	Max
All Products	Relevancy	2,342	4.48	0.88	1.00	4.50	7.00
	Surprise	2,342	3.76	0.85	1.00	3.75	7.00
Low Congruence	Relevancy	822	4.30	0.87	1.00	4.29	7.00
	Surprise	822	3.93	0.82	1.00	4.00	7.00
Medium Congruence	Relevancy	701	4.42	0.89	1.50	4.43	7.00
	Surprise	701	3.82	0.87	1.00	3.83	6.50
High Congruence	Relevancy	819	4.72	0.82	1.50	4.75	7.00
	Surprise	819	3.53	0.81	1.00	3.50	6.50

Note: The unit of observation is a product (i.e., audio-books and ebooks). The constructs are the average ratings from online participants (N=144).

Table 9: Mediation Analysis on Consumer Preference

	Model1		Model2	
	EST	SE	EST	SE
Intercept	4.134***	0.024	4.171***	0.019
Congruence	1.311***	0.154	0.079	0.116
Congruence Square	5.050***	1.069	0.239	0.790
Survey Construct			0.739***	0.016
Survey Construct Square			0.050***	0.013
$R^2$	0.031		0.458	
N	2,342		2,342	

Note: The unit of observation is products. The dependent variable is the extent to which the respondent is curious about consuming the content of the product, after viewing the product profile and the text description.

\*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ , + $p < 0.10$

## Image-Text Congruence Appendix

### Details on Image-Text Congruence


Three independent coders from a major US research university were invited to rate the level of congruence between an image-text pair. Figure A1 shows a screenshot of the study interface.

Figure A1: Image-Text Congruence Rating Interface for Human Annotators

ID: 000121

Jump To Pair 314

1 / 2000



Transportation Story: each vehicle takes its own roads! Transportation plays a very important role in people's everyday life, including both daily commute and travelling. How much do you know about transportation?

Based on the text summary and the cover image of a book shown above, how congruent do you think are the information conveyed from the book cover and the text description:(1 as "not congruent at all"; 10 as "absolutely congruent")

● 0 ● 1 ● 2 ● 3 ● 4 ● 5 ● 6 ● 7 ● 8 ● 9 ● 10

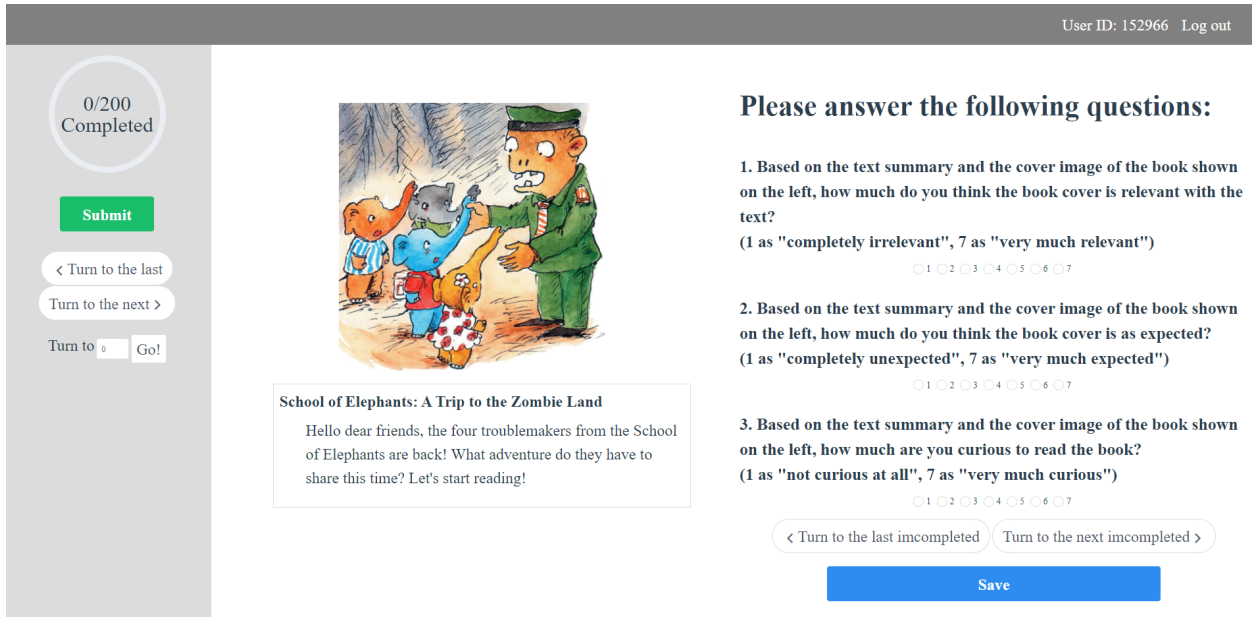
### Online Study Design Interface

Undergraduate students from a research university in Asia were invited to participate in our online study to explore mechanisms. Each student rated 200 pairs of image and text. An example pair is illustrated in Figure A2. For each pair, the student rated *relevancy* and *expectancy* using a 7-point Likert scale. The participant also rate how curious s/he was to read the book.

### Details on Model Comparison to Measure Congruence

We implement the whole pipeline of benchmark models with Python, using Jieba repository for Chinese word segmentation, Gensim repository for the LDA model, and Sklearn repository for linear regression. For this experiment, we use 1,800 image-text pairs as the training set and 200 as the test set.

Figure A2: Online Study Interface



To evaluate the performance of various methods, we use the human annotations as the ground truth and calculate the Pearson correlation with each of the congruence measures. To choose an appropriate topic number  $k$  and alleviate the randomness in LDA, we run the experiments 10 times with  $k$  in the range from 10 to 90. The results are shown in Figure A3, with the left panel corresponding to the unsupervised cosine method and the right panel showing the supervised linear regression method. The vertical axis is the average Pearson correlation (standard error) with human annotations, and the horizontal axis is the corresponding topic number. The highest correlation in the test set is around 0.35 for the Cosine method and around 0.56 for the regression method. Overall, the supervised model outperformed the unsupervised method for nearly the entire range of topic numbers.

## App Usage during the Day

Figure A4 depicts the distribution of the time of day when users become active on the app. Peak usage occurs after school between 8 p.m. and 10 p.m. During the day, usage peaks around noon and 1 p.m. The peak around noon is reasonable, because schools take a long lunch/after-lunch recess in many parts of China.

Figure A3: Performance of Benchmark Models

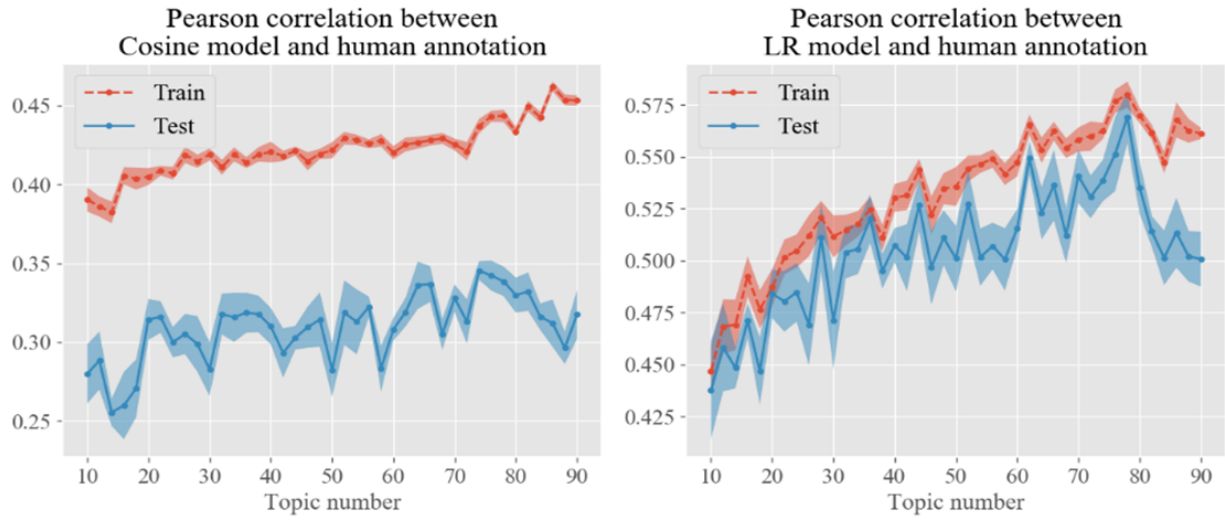


Figure A4: Time-of-Day Distribution for App Usage

