Electronic Commerce Research
https://doi.org/10.1007/s10660-025-10092-2

Multimodal topic models for brand competitive analysis

Xiaoning Wang[1] · Xiaoling Lu[2,3,4] · Yuxuan Guo[3] · Li'ao Zhu[3] · Feifei Wang[2,3,4]

Received: 22 July 2025 / Accepted: 25 December 2025

Abstract

Brand competitive analysis is crucial for businesses to understand their market position and develop strategies to outperform competitors. As consumer behavior and market trends evolve, companies increasingly rely on data-driven insights to stay ahead. In this context, online reviews on E-commerce platforms offer a wealth of information, providing valuable opportunities for analyzing consumer preferences. With the rise of multimodal reviews (combining both text and images), it has become essential to harness these diverse data sources effectively. To address this need, we first propose a multimodal and multi-corpora latent Dirichlet allocation (MM-LDA) model for static data. It uncovers both general topics and brand-specific topics, enabling fine-grained insights into consumer concerns and brand positioning. We further extend MM-LDA into a dynamic framework and propose the multimodal and multi-corpora dynamic topic model (MM-DTM), which incorporates temporal dynamics to capture topic evolution over time. The effectiveness of our approaches are demonstrated using two datasets from ZOL and Baidu Tieba discussions.

Keywords Brand Competitive Analysis · Dynamic Analysis · Multimodal Analysis · Social Media · Topic Models

X. Wang and X. Lu have contributed equally to this work.

✉ Feifei Wang
feifei.wang@ruc.edu.cn

[1] School of Data Science and Media Intelligence, Communication University of China, Beijing, China

[2] Center for Applied Statistics, Renmin University of China, Beijing, China

[3] School of Statistics, Renmin University of China, Beijing, China

[4] Innovation Platform, Renmin University of China, Beijing, China

Published online: 16 January 2026

1 Introduction

In today's competitive marketplace, brand competitive analysis is critical for businesses to assess their market position and craft strategies to outperform competitors. This analysis helps businesses identify their strengths and weaknesses, discover market opportunities, and refine strategies to excel in the face of competition. With the widespread use of social media and E-commerce platforms, businesses can now access vast amounts of consumer reviews on E-commerce platforms. These reviews provide detailed descriptions of products or services, reflecting performance, quality, and user experience. By analyzing these reviews, businesses can gain deep insights into how their products are perceived relative to competitors. Additionally, reviews highlight emerging trends and shifts in consumer preferences, enabling businesses to adjust their strategies proactively and maintain a competitive edge.

As consumers increasingly use images alongside text to express opinions, multimodal reviews are becoming an important source of information for businesses. Compared to traditional text reviews, multimodal reviews play an important role in enhancing user experience, influencing purchase decisions and promoting brand interaction and marketing [17, 23, 32]. For instance, Ma et al. [18] utilized deep learning methods to extract text and image features, demonstrating how image features enhance the accuracy of usefulness prediction. Li et al. [10] treated images as auxiliary variables and studied their impact on review helpfulness, showing that reviews with images are more helpful. These studies indicate that incorporating image information significantly improves the accuracy of text-based review mining tasks.

For brand competitive analysis, multimodal reviews are particularly valuable. The way consumers integrate textual and visual content in their reviews reflects a more holistic evaluation of brands. While textual descriptions often emphasize functional aspects, images typically highlight experiential or emotional factors such as aesthetics, usability, and satisfaction. Combining these perspectives enables a more comprehensive understanding of consumer perceptions, which is essential for assessing competitive positioning. Therefore, incorporating multimodal reviews provides deeper insights into brand differentiation and rivalry in the marketplace.

In the analysis of multimodal reviews, topic models have become a widely used approach. For example, Corr-LDA assumed a one-to-one correspondence between textual and visual topics [2], and subsequent extensions such as MMDF-LDA [35] integrated visual features of images with associated textual data to enhance annotation accuracy on social media. More recently, advanced models such as LB-MMT [12] and SMMTM [34] have been developed to capture the complex interplay between short texts and images in online platforms. These models have made important progress in bridging textual and visual modalities. However, existing studies on multimodal reviews are not well suited for brand competitive analysis. Most prior work has been limited to single corpora, lacking the capacity to jointly analyze multiple datasets, which is essential for comparative assessments across brands. In addition, these models are often static and thus fail to capture the temporal dynamics of consumer perceptions that are crucial for understanding the evolution of brand competitiveness. Furthermore, they often treat reviews as homogeneous data, with-

out explicitly accounting for brand-specific characteristics embedded in consumer opinions.

Previous studies have also applied topic models to the analysis of multiple corpora. For example, Wang et al. [27] introduced a latent Dirichlet allocation (LDA)-based approach to extract information from two corpora of competing products. Building on this work, Guo et al. [6] developed the Brand Joint LDA (BJ-LDA) model, specifically designed for mining multi-brand corpora. The BJ-LDA model can identify both general topics shared across brands and specific topics unique to each brand, offering deeper insights for brand competitive analysis. However, existing research on multi-corpora review mining primarily focuses on textual data, with relatively little attention given to multimodal content and the temporal dynamics inherent in consumer reviews.

To address the challenge of comparing multiple brands using multimodal reviews, we first develop a multimodal and multi-corpora latent Dirichlet allocation (MM-LDA) model for static review data. The MM-LDA model integrates both textual and visual information. It incorporates the scale-invariant feature transform (SIFT) image feature extraction and K-means clustering to analyze image comments alongside textual data. Additionally, it simultaneously processes multiple corpora to extract a range of topics, including both general topics shared across all brands and specific topics unique to individual brands. Each topic is represented through two distinct clusters: one for textual data and one for visual data. This dual-level analysis enables a deeper understanding of similarities and differences of brands. Furthermore, to handle dynamic multimodal reviews and capture the temporal evolution of topics, we further extend the model into a dynamic framework, resulting in the multimodal and multi-corpora dynamic topic model (MM-DTM). The effectiveness of the MM-LDA and MM-DTM models is demonstrated through extensive experiments on two datasets: the ZOL mobile phone dataset and the Baidu Tieba discussion dataset. These experiments illustrate the capability of the two models to reveal insightful general topics across all brands as well as specific topics for individual brands. These topics can further assist in analyzing competitive relationships among brands.

In summary, we provide the following contributions in this work. First, we provide a systematic framework that leverages both textual and visual reviews to assess brand competitiveness. Second, we propose the MM-LDA model, which identifies both general topics shared across brands and brand-specific topics, enabling detailed characterization of individual brands. Third, we extend MM-LDA to the MM-DTM model, enabling the capture of temporal dynamics in multimodal reviews and revealing how brand topics evolve over time. Last, we demonstrate the effectiveness of our proposed models through extensive experiments on real-world datasets.

The rest of the paper is organized as follows: Sect. 2 provides the literature review. Section 3 presents the proposed MM-LDA and MM-DTM. Section 4 presents experiments conducted on the ZOL dataset, while Sect. 5 covers experiments on the Baidu Tieba dataset. Finally, Sect. 6 concludes the paper with a brief discussion.

2 Related work

2.1 Research of brand competitive analysis

Brand competitive analysis can be approached from various perspectives. In this section, we focus on the analysis based on consumer reviews. From this angle, brand comparison can be conducted using multi-corpora analysis methods, which is a crucial research area with significant implications for understanding and comparing multiple datasets. One key approach within this field is comparative summarization methods, which are particularly effective for analyzing documents on similar topics and identifying differences between them. These methods support decision-making and provide valuable insights for multi-corpora analysis. For instance, the linear programming method [8] and discriminatory sentence recognition method [25] have been proposed for comparative mining in purely textual data. In the context of online reviews, Sipos and Joachims [22] introduced an innovative objective function to pair review fragments from multiple corpora, enabling the identification of sentences that compare similar features. Latent Dirichlet allocation (LDA) is a prominent method for extracting thematic information from text, making it a key tool for text mining [3]. Several models based on LDA have been developed for multi-corpora analysis, which can be easily extended for brand competitive analysis. For example, Wang et al. [27] proposed an LDA-based model to analyze two competing brands from a topic perspective. This model effectively analyzed two separate corpora and leveraged review data to its full extent. Building on this, Guo et al. [6] introduced a brand-focused LDA model, mining both shared and unique topics among different brands to facilitate comprehensive brand analysis. This approach allows for comparing the strengths and weaknesses of several brands.

Compared with existing studies on brand competitive analysis, our approach offers two notable advantages. First, it incorporates image information in addition to textual reviews, enabling a more comprehensive understanding of consumer perceptions and brand attributes. Second, by extending the model into a dynamic framework, it captures the temporal evolution of topics, thereby allowing the analysis of how brand competition and consumer concerns change over time.

2.2 Research on multimodal learning

Multimodal representation methods are an actively researched area, primarily divided into two categories: traditional multimodal representation methods and deep learning approaches. In traditional multimodal representation techniques, Li et al. [11] differentiated categories of linguistic content to encode textual information, then utilized open tools to extract interpretable features such as color and image quality, and encoded images. Xia et al. [28] emphasized enhancing audio and visual features under the guidance of textual semantics, showcasing a method where text modality's global semantics improve the representation learning of audio and visual encoders.

Within the deep learning framework for multimodal representation, Lu et al. [15] proposed a dual-stream architecture for processing visual and linguistic inputs separately, with shared transformer layers for fusion. This design facilitates effec-

tive cross-modal interactions, significantly improving performance across various vision-and-language tasks. Radford et al. [20] proposed contrastive language–image pre-training (CLIP) to demonstrate how pre-training on large-scale image-text pairs yielded a versatile visual-linguistic representation capable of zero-shot or few-shot learning across multiple visual tasks. CLIP's success highlights the potential of large-scale cross-modal data pre-training. Zhai et al. [33] addressed the challenge of named entity recognition in social media texts by incorporating visual information along with textual content. However, deep learning-based approaches to multimodal representation typically suffer from poor interpretability and a high dependency on large-scale training datasets.

In the study of topic models for multimodal data, Blei and Jordan [2] proposed the multimodal topic model, which assumed a one-to-one correspondence between the topics of textual and image modalities. Putthividhya et al. [19] presented an image annotation method that combined topic regression with multimodal LDA, utilizing LDA to analyze image content and integrating topic regression techniques to enhance the accuracy and relevance of image annotations. Zheng et al. [35] proposed an improved multimodal LDA model (MMDF-LDA) for annotating images on social media. This model integrated the visual features of images with associated textual data, using LDA to identify and label key topics and concepts within images, thereby improving the precision and efficiency of annotating social media content. Zhang et al. [34] introduced the social media multimodal topic model (SMMTM) to specifically address the intricacies of short texts within social media, using the SIFT algorithm for processing images. This model operates on the assumption that a given document's text is aligned with a single topic, whereas its associated images might pertain to multiple topics. Li et al. [12] processed images using the SIFT algorithm, transforming them into visual words, then combined this with textual data for LDA topic modeling.

Compared with prior studies in multimodal learning, our models exhibit several distinctive features. First, it incorporates multimodal information into topic modeling across multiple corpora and further extends to a dynamic setting, enabling the analysis of topic evolution over time. Second, our proposed models are capable of identifying both general topics shared across brands and brand-specific topics, thus providing a fine-grained representation of brand-level distinctions.

3 Methodology

3.1 Multimodal multi-corpora LDA(MM-LDA)

To fully leverage the rich and diverse information from multiple multimodal corpora for brand competitive analysis, we introduce the MM-LDA model in this section. As illustrated in Fig. 1, the MM-LDA model comprises two main steps. Since the review data analyzed by MM-LDA includes both text and images, the first step involves aligning these two modalities for effective analysis. For image data, we utilize the SIFT method [14] and K-means clustering to convert images into visual Bag-of-Words (BoW) representations. For textual data, it is straightforward to transform

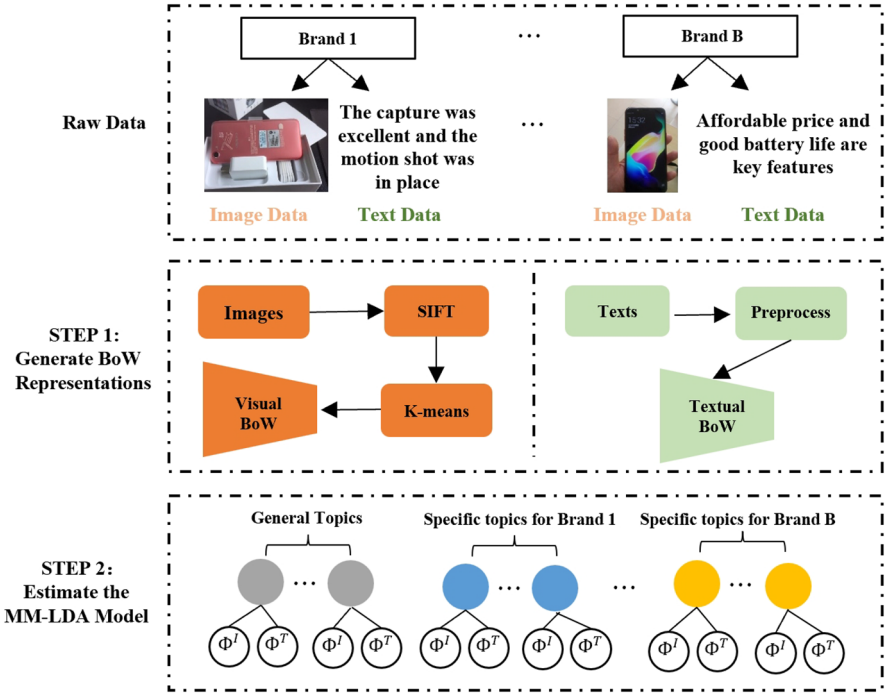


Fig. 1 The MM-LDA Framework

it into a BoW format. The BoW representations of both image and text modalities are then integrated and analyzed in the second step, which involves estimating the MM-LDA model. Step 2 of Fig. 1 provides a brief topic structure of MM-LDA, with a detailed construction presented in Fig. 3. This model is designed to uncover both general topics that are common across multiple corpora and specific topics unique to each corpus. By analyzing these general and specific topics, the model facilitates a deeper understanding of the relationships and differences among the multiple corpora.

To jointly exploit both textual and visual information, we assume that each topic is characterized by two distinct clusters of words: one for textual content (the corresponding distribution is Φ^T) and one for visual content (the corresponding distribution is Φ^I). This design is motivated by methodological considerations. Specifically, text and images represent heterogeneous modalities: textual reviews primarily capture functional attributes, whereas images often highlight experiential or emotional aspects such as aesthetics and usability. Treating them as a single cluster would obscure these complementary dimensions. By modeling separate distributions, our framework is able to preserve modality-specific semantics while still linking them under the same latent topic. This design is also consistent with previous research on multimodal topic models [2, 12, 21, 30], where distinct distributions are assigned to different modalities to capture their unique characteristics.

In the following sections, we will first describe the process of generating BoW representations for images and texts (Step 1). Once we have obtained the BoW representations, we will proceed to establish the MM-LDA model (Step 2). The model

structure of MM-LDA will be detailed in Section 3.1.2, and the detailed estimation method is left in Appendix A.

3.1.1 Generating BoW representations

Before estimating the MM-LDA model, we need to generate BoW representations for both text and image data. For textual data, the process of generating BoW is relatively straightforward. Typically, we remove stop words, filter out low-frequency words, and eliminate numbers. This approach is generally sufficient for text. However, for image data, the process is more complex. Thus in the following, we will focus on the process of generating BoW representations for images. Here we basically apply the bag-of-visual-words (BoVW) method to process image information, which is a classic approach for image representation and widely used in image retrieval and classification [9, 31]. Unlike generating BoW representations for texts, BoVW encodes images based on the statistical frequency of visual words and involves three main steps: (1) feature extraction using the SIFT method, (2) K-means clustering, and (3) statistical distribution representation. First, the SIFT algorithm [14] identifies keypoints in images and extracts 128-dimensional feature vectors from these points. We use the *opencv* library in Python to extract up to 100 keypoints per image. Figure 2 shows the keypoints of one image, for example, on the ZOL smartphone dataset, which we analyze in Sect. 4. It demonstrates that, the SIFT method can effectively identify significant features from images.

Next, the SIFT features are processed using the K-means clustering method to construct a visual vocabulary, with each cluster centroid representing a visual word. Assume the number of clusters to be $\widetilde{K}$, which can be determined by monitoring the performance of the subsequent topic modeling. Then each image is represented by $\widetilde{K}$ visual words. These visual words are aggregated to form a visual word list for the image modality of each review, facilitating subsequent topic modeling analysis.

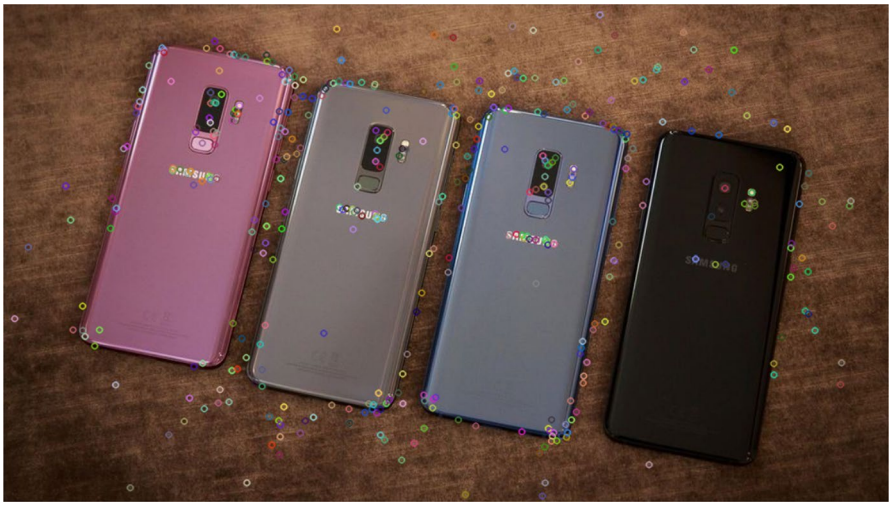


Fig. 2 Illustration for SIFT Extraction

Finally, we generate a frequency histogram to represent the distribution of visual words within each image. If the most frequent visual word in an image exceeds the second most frequent by 10%, then the image would be categorized by the dominant visual word. Accordingly, we also consider images that could not be assigned a dominant visual word as *uninformative*. In the ZOL dataset, we find about 20% of images fall into this category. Specifically, the proportions of uninformative images for Huawei, Oppo, Samsung, and Vivo are 20.00%, 19.27%, 21.74%, and 19.14%, respectively. These results indicate that the majority of images provide meaningful visual information. Table 1 presents some examples of the computed visual words in the ZOL dataset. It is obvious that, images classified under the same visual word category exhibit similar meanings, validating the effectiveness of this workflow in extracting interpretable visual vocabulary for multimodal topic modeling.

In addition to the SIFT-based approach, we also investigate the performance of using modern deep learning–based visual embedding methods. Specifically, we experiment with a more advanced approach, namely CLIP-based visual features [4,

Table 1 Examples of Four Dominant Visual Words, Their Practical Illustrations, and Representative Images in the ZOL Dataset

Visual Word 1	Practical illustration: system configuration
	
Visual Word 2	**Practical illustration: phone screen display**
	
Visual Word 3	**Practical illustration: the taken photos**
	
Visual Word 4	**Practical illustration: phone appearance design**
	

13], and compare its performance with that of the SIFT-based model. The detailed results are provided in Appendix D. We find that the SIFT-based method yields comparable or even slightly better results than the CLIP-based method, suggesting that the proposed framework is relatively robust to different choices of image representation.

3.1.2 The generative process of MM-LDA

MM-LDA aims to extract both general and specific topics from multiple corpora, each containing two modalities: textual and image data. First, we introduce some necessary notations. Suppose one product category involves B different brands. For brand b, the number of documents is defined as D_b, where b ranges from 1 to B. The textual vocabulary size for all texts from brand b is denoted as V_b^T. Similarly, the visual vocabulary size from the visual BoW data extracted from images is denoted as V_b^I. Summarizing the corpora of all brands, the total textual vocabulary size is V^T, the total visual vocabulary size is V^I, and the total number of documents is D.

Furthermore, we assume the existence of general topics shared across all brands, as well as specific topics corresponding to each brand. Specifically, we assume there are K_0 general topics across all D review documents. Each general topic corresponds to two clusters: one for textual words and one for image visual words. This model structure aids in exploring common interests across all brands. Additionally, for each brand b, where $b \in \{1, 2, \ldots, B\}$, we assume there are K_b specific topics, with different brands having different numbers of specific topics. These K_b specific topics consist of K_b clusters of textual words specific to the brand and K_b clusters of image visual words specific to the brand. Other notations are summarized in Table 2.

Based on the above notations, we present the generative process of MM-LDA, which is summarized in Fig. 3. We first draw word distributions of each general topic from a Dirichlet prior with hyperparameters β. For each general topic $k \in \{1, ..., K_0\}$, there are two types of words distributions: one on textual vocabulary, which is denoted by $\Phi_k^{T,g}$, and the other on image visual vocabulary, which is denoted by $\Phi_k^{I,g}$. Then for each brand b, we draw word distributions of its K_b specific topics from a Dirichlet prior with hyperparameters β_b. There are also two types of word distributions on each specific topic $k_b \in \{1, ..., K_b\}$: one on textual vocabulary, which is denoted by $\Phi_{k_b}^{T,s}$, and the other on image visual vocabulary, which is denoted by $\Phi_{k_b}^{I,s}$.

Next, similar to the standard LDA model, for each review document d in brand b, we draw two document-topic distributions: one general document-topic distribution $\theta_{d_b}^g \sim \mathrm{Dir}(\alpha_g)$ and one specific document-topic distribution $\theta_{d_b}^s \sim \mathrm{Dir}(\alpha_s)$, where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution and α_g and α_s are hyperparameters. Next, we draw topic indicators $z_{d_b,n}^g \sim \mathrm{Multi}(\theta_{d_b}^g)$ and $z_{d_b,n}^s \sim \mathrm{Multi}(\theta_{d_b}^s)$ for each textual word $w_{d_b,n}^T$ and image visual word $w_{d_b,n}^I$ in document d. The n-th textual word $w_{d_b,n}^T$ in document d can describe either a general word or a specific word. It is similar for the n-th image visual word $w_{d_b,n}^I$ in document d. Thus we introduce two indicator variables $u_{d_b,n}^T$ and $u_{d_b,n}^I$ to distinguish between the types of textual word $w_{d_b,n}^T$ and image visual word $w_{d_b,n}^I$. Take the textual modality as an example. Let $u_{d_b,n}^T$ fol-

Table 2 Notations in MM-LDA

Symbol	Meaning	Symbol	Meaning
B	Number of brands	K_0	Number of general topics
K_b	Number of specific topics in brand b	β	Dirichlet prior
D	Total number of documents	D_b	Number of documents in brand b
V^T	Total number of textual words	V_b^T	Number of specific textual words in brand b
V^I	Total number of image visual words	V_b^I	Number of specific image visual words in brand b
p	Parameter of Bernoulli distribution	γ	Parameter of Beta distribution
$N_{d,n}$	Word in document	α	Dirichlet prior
$\Phi_k^{T,g}$	General textual word distribution	$\Phi_k^{I,g}$	General image visual word distribution
$\Phi_{k_b}^{T,s}$	Specific textual word distribution	$\Phi_{k_b}^{I,s}$	Specific image visual word distribution
$\theta_{d_b}^g$	General topic distribution	$\theta_{d_b}^s$	Specific topic distribution
$Z_{d_b,n}^g$	Indicator of general topic of word	$Z_{d_b,n}^s$	Indicator of specific topic of word
$w_{d_b,n}^I$	Observed image visual word	$u_{d_b,n}^I$	Indicator of general/specific image visual word
$w_{d_b,n}^T$	Observed textual word	$u_{d_b,n}^T$	Indicator of general/specific textual word

lows the Bernoulli distribution with parameter p, where p follows the Beta distribution with hyperparameter γ. Then the extraction of $w_{d_b,n}^T$ and $w_{d_b,n}^I$ is generated as follows:

$$w_{d_b,n}^T \sim \begin{cases} \text{Multi}(\Phi_{z_{d_b,n}^g}^{T,g}) & \text{if } u_{d_b,n}^T = 0 \\ \text{Multi}(\Phi_{z_{d_b,n}^s}^{T,s}) & \text{if } u_{d_b,n}^T = 1 \end{cases}$$

$$w_{d_b,n}^I \sim \begin{cases} \text{Multi}(\Phi_{z_{d_b,n}^g}^{I,g}) & \text{if } u_{d_b,n}^I = 0 \\ \text{Multi}(\Phi_{z_{d_b,n}^s}^{I,s}) & \text{if } u_{d_b,n}^I = 1 \end{cases}$$

Step 1: Generate word distributions for topics
for the general topic $k \in \{1, 2, 3, \cdots, K_0\}$ **do**
  Generate its textual-word distribution $\phi_k^{T,g} \sim Dir(\beta)$
  Generate its general visual-word distribution $\phi_k^{I,g} \sim Dir(\beta)$
end for
for the corpus $b \in \{1, 2, \cdots, B\}$ **do**
  for the specific topic $k_b \in \{1, 2, \cdots, K_b\}$ **do**
    Generate its textual-word distribution $\phi_{k_b}^{T,s} \sim Dir(\beta_b)$
    Generate its visual-word distribution $\phi_{k_b}^{I,s} \sim Dir(\beta_b)$
  end for
end for
Step 2: Generate topic distributions for documents
for the corpus $b \in \{1, 2, \cdots, B\}$ **do**
  for the document d in corpus b **do**
    Generate its general topic distribution $\theta_d^g \sim Dir(\alpha_g)$
    Generate its specific topic distribution $\theta_{d_b}^s \sim Dir(\alpha_s)$
  end for
end for
Step 3: Generate observed textual and image visual words
for the textual word $w_{d_b,n}^T$ in each document **do**
  Sample its general topic indicator $z_{d,n}^g \sim Multi(\theta_d^g)$
  Sample its specific topic indicator $z_{d_b,n}^s \sim Multi(\theta_{d_b}^s)$
  Sample the general-or-specific indicator $u_{d_b,n}^T \sim Bernoulli(p)$ with $p \sim Beta(\gamma)$
  if $u_{d_b,n}^T = 0$ (describing the general topic) **then** $w_{d_b,n}^T \sim Multi(\phi_{z_{d,n}}^{T,g})$
  else $w_{d_b,n}^T \sim Multi(\phi_{z_{d_b,n}}^{T,s})$
  end if
end for
for the visual word $w_{d_b,n}^I$ in document **do**
  Sample its general topic indicator $z_{d,n}^g \sim Multi(\theta_d^g)$
  Sample its specific topic indicator $z_{d_b,n}^s \sim Multi(\theta_{d_b}^s)$
  Sample the general-or-specific indicator $u_{d_b,n}^I \sim Bernoulli(p)$ with $p \sim Beta(\gamma)$
  if $u_{d_b,n}^I = 0$ (describing the general topic) **then** $w_{d_b,n}^I \sim Multi(\phi_{z_{d,n}}^{I,g})$
  else $w_{d_b,n}^I \sim Multi(\phi_{z_{d_b,n}}^{I,s})$
  end if
end for

Algorithm 1 Generation process of MM-LDA

The Bernoulli/Beta switch in our model is designed to explicitly distinguish between general and brand-specific topics in a simple and interpretable manner. The Bernoulli variable $u_{d_b,n}^T$ directly indicates whether a topic belongs to the general or brand-specific category, while the Beta prior allows flexible modeling of the probability of the Bernoulli indicators. This design is also supported by previous studies on topic modeling, which have employed similar switches to separate different topic components; see Guo et al. [6], Lu et al. [16], Wang et al. [26] for examples. Except for this design, using hierarchical priors is another option. However, compared to hierarchical priors, this Bernoulli/Beta approach maintains conceptual clarity, reduces model complexity, and facilitates efficient Gibbs sampling inference. Investigation of hierarchical priors serves as an interesting direction for future work. The generation process of the MM-LDA model can be summarized in Algorithm 1. The detailed estimation procedure of MM-LDA can be found in Appendix A.

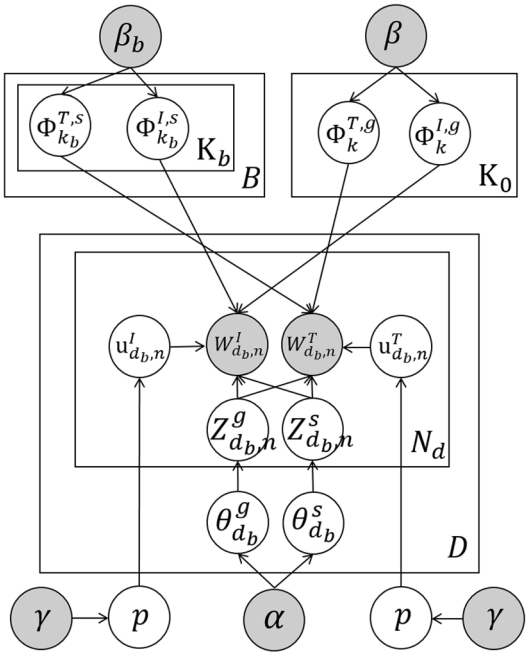


Fig. 3 The Generation Process of MM-LDA

3.2 Multimodal multi-corpora DTM (MM-DTM)

In real-world scenarios, consumer reviews are not only multimodal but also dynamic, evolving over time as brand perceptions shift, new products are launched, and marketing strategies change. Temporal patterns in consumer reviews, such as emerging concerns and shifting preferences, can carry valuable signals for understanding market dynamics. These dynamic characteristics can significantly enhance the analysis of brand competition by revealing how topics evolve and how different brands respond to consumer feedback over time. To capture such temporal dynamics, we extend our MM-LDA model to a dynamic version (referred to as MM-DTM), which incorporates the temporal dimension of multimodal consumer reviews. To enable the MM-DTM model to capture the temporal evolution of topics, the original review corpus is divided into multiple time slices based on timestamps.

In MM-DTM, we denote both textual and visual words uniformly as w. Accordingly, $Z_{db,w}$ and $U_{db,w}$ represent the topic assignment and indicator for either textual or visual words, depending on the context. When necessary, $w \in W^T$ implies a textual word; $w \in W^I$ implies a visual word. Recall there are B brands in the whole dataset. At time t with $1 \leq t \leq T$, we collect a total of $D_{b,t}$ documents for brand b, whose corpus is represented by $\mathbb{D}_{b,t}$. For each document $d \in \mathbb{D}_{b,t}$, it contains two kinds of information, i.e., $d = \{\boldsymbol{w}_{d,t}, \boldsymbol{v}_{d,t}\}$. Here $\boldsymbol{w}_{d,t} = \left\{w^t_{d,1}, w^t_{d,2}, \ldots, w^t_{d,N_d}\right\}$ denote a set of textual words in document d from a textual vocabulary $\mathbb{V}_w$, where $w^t_{d,i}$ is the i-th word and N_d denotes the total number of words in document d. Let $\boldsymbol{v}_{d,t} = \left\{v^t_{d,1}, v^t_{d,2}, \ldots, v^t_{d,C_d}\right\}$ denote visual words in document d from a visual

vocabulary $\mathbb{V}_v$, where $v^t_{d,i}$ is the *i*-th visual word in the document *d* and C_d denote the total number of visual words in document *d*. For each textual word *w*, we define U_w as a Bernoulli indicator to determine whether it represents a general ($U_w = 1$) or brand-specific topic ($U_w = 0$). Here *p* is the probability of the Bernoulli distribution, which is further generated from a Beta distribution with the parameter γ. For each visual word *v*, a similar indicator U_v is defined.

Assume there are two types of topics underlying the whole corpus. Specifically, there are K_0 general topics and K_b brand-specific topics with $1 \le b \le B$. For each general topic $1 \le k \le K_0$, we define $\beta^g_{k,t}$ as a $\mathbb{V}_w$-dimensional multi-normal distribution over the textual words and define $\phi^g_{k,t}$ as a $\mathbb{V}_v$-dimensional multi-normal distribution over the visual words. Similarly, for brand-specific topic $1 \le k \le K_b$, define $\beta^s_{k,t}$ and $\phi^s_{k,t}$ to be the textual word or visual word distribution. Furthermore, for each document $d \in \mathbb{D}_{b,t}$, let θ^g_d define the document-level distribution over K_0 general topics and let θ^s_d define the document-level distribution over K_b brand-specific topics.

Based on the above notations, the generation process of MM-DTM can be summarized in Algorithm 2. The graphical representation of MM-DTM is present in Fig. 4. The estimation details of MM-DTM are present in Appendix B.

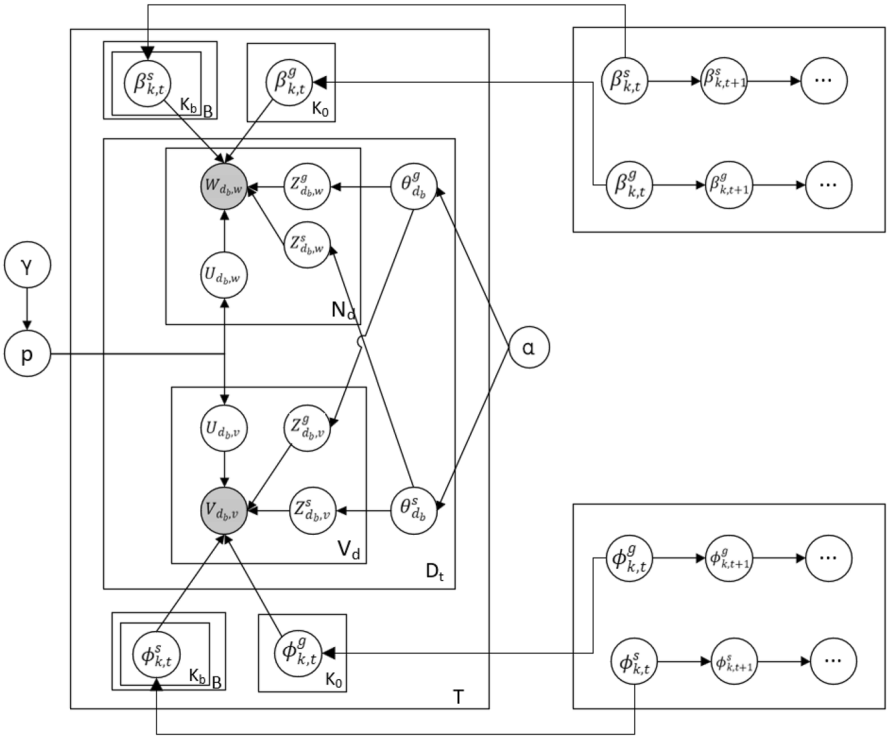


Fig. 4 The generative process of MM-DTM

Step 1: Evolve topic-word distributions over time
for each general topic $k = 1, \ldots, K_0$ **do**
  $\beta^g_{k,t} \sim \mathcal{N}(\beta^g_{k,t-1}, \sigma^2 I_w)$
  $\phi^g_{k,t} \sim \mathcal{N}(\phi^g_{k,t-1}, \sigma^2 I_v)$
end for
for each brand-specific topic $k = 1, \ldots, K_b$, brand $b = 1, \ldots, B$ **do**
  $\beta^s_{k,t} \sim \mathcal{N}(\beta^s_{k,t-1}, \sigma^2 I_{\mathbb{V}_w})$
  $\phi^s_{k,t} \sim \mathcal{N}(\phi^s_{k,t-1}, \sigma^2 I_{\mathbb{V}_v})$
end for
Step 2: Sample indicator prior
$p \sim \text{Beta}(\gamma)$
Step 3: Generate documents at each time t
for each brand $b = 1, \ldots, B$ **do**
  for each document $d \in \mathbb{D}_{b,t}$ **do**
    $\theta^g_d \sim \text{Dir}(\alpha), \quad \theta^s_{d_b} \sim \text{Dir}(\alpha)$
    for each textual word w **do**
      $U_{d_b,w} \sim \text{Bernoulli}(p)$
      if $U_{d_b,w} = 1$ **then**
        $Z^g_{d_b,w} \sim \text{Multi}(\theta^g_{d_b}), \quad z \leftarrow Z^g_{d_b,w}$
        $W_{d_b,w} \sim \text{Multi}(\pi(\beta^g_{z,t}))$
      else
        $Z^s_{d_b,w} \sim \text{Multi}(\theta^s_{d_b}), \quad z \leftarrow Z^s_{d_b,w}$
        $W_{d_b,w} \sim \text{Multi}(\pi(\beta^s_{z,t}))$
      end if
    end for
    for each visual word v **do**
      $U_{d_b,v} \sim \text{Bernoulli}(p)$
      if $U_{d_b,v} = 1$ **then**
        $Z^g_{d_b,v} \sim \text{Multi}(\theta^g_{d_b}), \quad z \leftarrow Z^g_{d_b,v}$
        $V_{d_b,v} \sim \text{Multi}(\pi(\phi^g_{z,t}))$
      else
        $Z^s_{d_b,v} \sim \text{Multi}(\theta^s_{d_b}), \quad z \leftarrow Z^s_{d_b,v}$
        $V_{d_b,v} \sim \text{Multi}(\pi(\phi^s_{z,t}))$
      end if
    end for
  end for
end for

Algorithm 2 Generation process of MM-DTM

4 Experiments on ZOL

4.1 Data description

In this section, we assess the efficacy of our proposed MM-LDA model on the ZOL mobile phone dataset [29]. Mobile phones are a significant consumer category, and consumers invest considerable time in making purchasing decisions, often leaving detailed comments on E-commerce platforms. Consequently, the opinions expressed in these online comments are rich and varied. For our analysis, we selected four well-known mobile phone brands in China based on brand characteristics and price. They

are, respectively, Huawei, Samsung, OPPO, and VIVO. According to their official websites, these brands represent two high-end innovative brands and two high cost-performance brands, each with specific product positioning. For instance, Huawei focuses on high-end innovation with strengths in mobile phone systems, while OPPO targets the youthful market and is popular among young users for its excellent design and photography features.

Consumer reviews, being unstructured information, often contain a significant amount of noise. Therefore, data preprocessing is a crucial step in review mining. Our preprocessing involved several steps. The first step was filtering, where we removed duplicate comments with the same buyer ID or identical content. Additionally, we excluded comments that contained only numbers or punctuation. The second step was word segmentation, which we performed using the “Jieba" tool in Python. The third step was the removal of stop words, for which we used an extended version of the standard stop word list. Table 3 provides descriptive statistical information about the ZOL dataset after preprocessing.

4.2 Analysis using the MM-LDA model

To apply the MM-LDA model, we first need to generate the BoW representations for both images and texts. For textual data, we remove stop words, eliminate numbers, and filter out words whose frequency is lower than 10%. For images, we first use the SIFT method to extract features and then apply the K-means clustering method with $\widetilde{K} = 50$. Other $\widetilde{K}$ values are also considered, but the corresponding topic coherence and model perplexity metrics under the same number of topics are suboptimal.

After generating the BoW representations, we try to estimate the MM-LDA model. The hyperparameters α, γ, and the βs need to be specified in advance. Following prior studies on topic modeling (e.g., Griffiths and Steyvers [5], Wang et al. [24], Guo et al. [6]), we set $\alpha = 0.2$ and all βs to be 0.1. These values have been widely adopted as reasonable defaults, leading to interpretable and stable topic distributions. As for the hyperparameter γ, we set it to 0.5, since it serves as a Beta prior for the probability of selecting a general or brand-specific topic. A symmetric prior with $\gamma = 0.5$ provides a noninformative yet slightly sparse preference, allowing the model

Table 3 The Detailed Information of Four Mobile Phone Brands

Brand	Price (rmb)	Brand Features	Review Count	Review Length	Image Count
Huawei	4000	Focusing on high-end innovation	374	285	3.67
Samsung	4500	Positioning in the mid-to-high-end market	472	226	3.81
VIVO	2000	High cost-performance ratio	324	312	3.27
OPPO	2000	Emphasizing appearance and photography	358	241	4.02

to balance the two types of topics without biasing toward either side. To assess the impact of these hyperparameters, we further conduct sensitivity analyses to examine model stability under different parameter settings; see Appendix E for details. The results show that our model performance remains robust across a reasonable range of hyperparameter values, supporting the appropriateness of our parameter calibration.

Except for the hyperparameters, we also need to determine the general topic number K_0 and specific topic number K_b for each corpus. To this end, we initially set $K_0 = 1$ and $K_b = 1$ for each brand (referenced as Model 1). Then we increase K_0 and K_b by a stepsize of 1, with the maximum value set at 6. In total, we experiment with 15 different candidate models, and the specific number of general topics and specific topics for each model can be found in Table 13 in Appendix C. During this process, we compute the model perplexity (MP) and topic coherence scores (TC) to evaluate performance of MM-LDA under different topic numbers. Figure 5 presents the results of model perplexity and topic coherence scores. As shown, we find the model perplexity decreases to a stable level, and the topic coherence reaches a relatively high level. As a result, the optimal selection is Model 14. Correspondingly, the optimal number of general topics is $K_0 = 5$ and the optimal number of specific topics for Huawei, VIVO, Samsung, and OPPO are 4, 3, 4, and 2, respectively.

To complement the quantitative metrics of perplexity and topic coherence, we further conduct a human evaluation to assess topic quality. Specifically, three annotators with backgrounds in marketing and data analysis independently rate the interpretability and coherence of the top topics generated by each model on a 5-point Likert scale. The average inter-rater agreement, measured by Cohen's κ, exceeds 0.75, indicating a high level of consistency among annotators. The results of this evaluation corroborate our selection of Model 14 as the optimal choice.

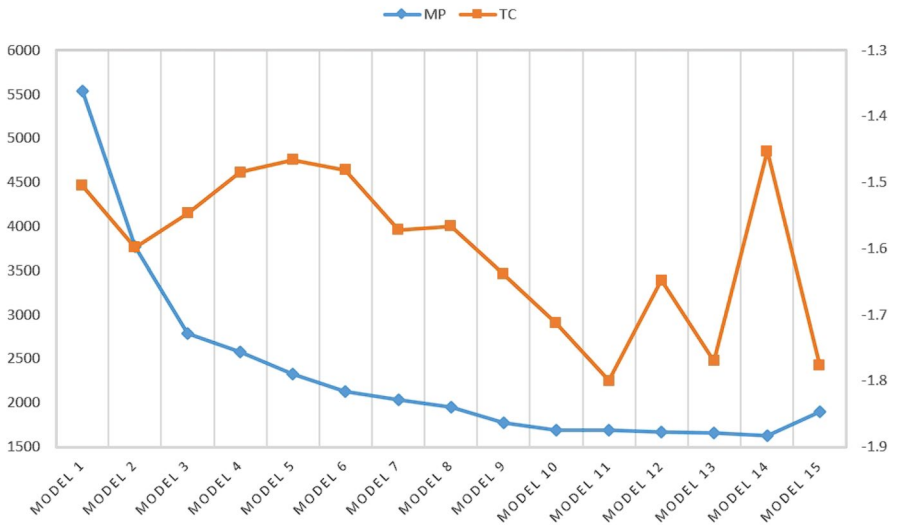


Fig. 5 The model perplexity (MP) and topic coherence scores (TC) during the model selection process for the ZOL data, where model 14 is the optimal choice

To evaluate the effectiveness of the proposed MM-LDA model, we compare it against several baselines, including the basic LDA model [1], the Dynamic Topic Model (DTM) [1], the BJ-LDA model for multi-corpus analysis [6], and the neural multimodal topic model Multimodal-ZeroShotTM [4]. For LDA, DTM, and Multimodal-ZeroShotTM, which are all designed for single-corpus modeling, we first merge the multiple brand corpora and then perform topic modeling. The number of topics for the merged corpus is set to 5. In contrast, the BJ-LDA model supports multi-corpus analysis but relies solely on textual data. For a fair comparison, the number of topics in BJ-LDA is set to be consistent with that in MM-LDA. We quantitatively compare these models using topic coherence and perplexity as evaluation metrics. The results, summarized in Table 4, demonstrate that the proposed MM-LDA model achieves better performance than all baselines.

Last, we quantify the correspondence between visual and textual high-probability words for each topic. To this end, we first assign each review to the topic with the highest probability in its inferred topic distribution θ. For each topic k, we identify the top-N textual words and top-N visual words based on $\Phi_k^{T,\cdot}$ and $\Phi_k^{I,\cdot}$, respectively. We then compute the co-occurrence probability for each textual-visual word pair (w_i, v_j) as the fraction of reviews assigned to topic k, in which both w_i and v_j appear. Finally, we summarize these co-occurrence values into a single alignment score for the topic by averaging over all top-N word pairs. A higher score indicates stronger correspondence between the visual and textual words within a topic. The results show that for MM-LDA, the average score across topics is 51.64%. For comparison, we modify the classic LDA model by incorporating image features and assuming that each topic has distributions over two sets of words (i.e., textual and visual). We denote this model as multi-LDA, for which the average score is only 35.3%. These results suggest that, in MM-LDA, the visual and textual words within each topic exhibit a stronger and more coherent degree of semantic alignment.

4.3 Ablation study

To evaluate the contribution of multimodal inputs and the general/specific topic decomposition, we conduct a series of ablation experiments based on the proposed MM-LDA framework. Four variants are implemented for comparison:

(1) MM-LDA (Full model): The complete multimodal topic model integrating both textual and visual features, with general and brand-specific topics jointly modeled.
(2) Text-only model: Visual features are removed, and topic inference relies solely on textual reviews while keeping the same parameter settings as MM-LDA.

Table 4 Results of coherence and perplexity of different models on the ZOL dataset

Model	Coherence	Perplexity
LDA	−1.766	1602.641
BJ-LDA	−1.674	2415.586
Multimodal-ZeroShotTM	−11.512	15521.933
MM-LDA	−1.453	1180.975

(3) Image-only model: Textual features are excluded, and the model learns topics purely from image representations under identical configurations.
(4) General-topics-only model: Only the shared topics across brands are retained, while the specific topics are removed.

The quantitative results are summarized in Table 5, showing both topic coherence and perplexity metrics. Several key patterns can be observed. When only textual or only visual information is used, the model's perplexity increases sharply, indicating a substantial degradation in modeling quality and confirming the complementary nature of multimodal signals. In particular, the text-only model exhibits the highest perplexity and the lowest coherence, implying that textual content alone cannot sufficiently capture the semantic diversity of multimodal data. The image-only model performs slightly better than the full model in coherence, as visual features often cluster around salient product attributes, producing more homogeneous topics. Moreover, when we retain only the general topics and discard brand-specific ones, the total number of topics decreases, yielding lower topic coherence but improved perplexity. This suggests that while the simplified model fits the data more tightly, it loses interpretability and fails to represent fine-grained brand-level distinctions. Overall, these findings demonstrate that both multimodal integration and the incorporation of brand-specific topics are critical for balancing topic coherence and perplexity, thereby ensuring a more interpretable and semantically rich representation of multimodal content.

Lastly, note that the ZOL dataset contains a relatively small number of reviews for each brand. To investigate the robustness of MM-LDA, we conduct multiple runs with different random seeds on the ZOL data. Results show that the topic coherence and perplexity are consistent across runs. Furthermore, to validate MM-LDA on larger-scale data, we apply MM-LDA and the baseline models to the Baidu Tieba dataset used in Sect. 5, treating it as a static dataset, since it has a larger sample size. The results demonstrate that MM-LDA continues to outperform the baselines, confirming its effectiveness and generalizability to larger corpora. Please see Appendix F for details.

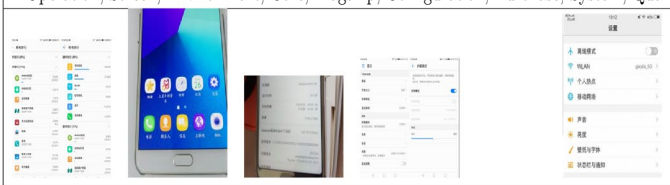
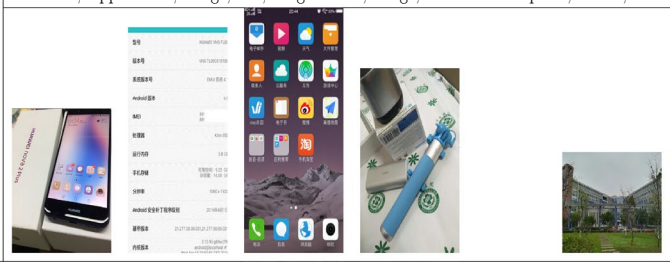
4.4 Brand competitive analysis using MM-LDA

The ZOL dataset consists of five general topics, each accompanied by corresponding high-frequency textual words and images, as illustrated in Table 6. As shown, these images visually depict the high-frequency words associated with general topics of common interest to users across all mobile phone brands. The first topic addresses mobile operating system configurations, featuring high-frequency words such as "operation", "environment", "configuration", and "system". The associated images primarily consist of screenshots related to mobile system settings. The sec-

Table 5 Results of ablation experiments on MM-LDA and Its variants

Model	Coherence	Perplexity
MM-LDA (Full model)	−1.453	1180.975
General-topics-only	−1.514	219.264
Text-only	−3.163	46617.193
Image-only	−0.775	7300.193

Table 6 Results of General Topics for the ZOL Data

General Topic 1	Operation, Screen, Environment, Core, Flagship, Configuration, Purchase, System, Quality, mAh
	
General Topic 2	Configuration, Fast charge, Quality, mAh, Release, NFC, Image, Suitable, Flagship, Customization
	
General Topic 3	Quality, Appearance, Large screen, Purchase, Placement, Image, Blur, Color
	
General Topic 4	Photography, System, Design, Like, Configuration, Experience, Effect, Good, Camera, Function
	
General Topic 5	Purchase, Appearance, Image, HD, Large screen, Usage, Power consumption, Sound, Screen, Speaker
	

ond topic focuses on mobile battery and configurations, including high-frequency words like “fast charge", “quality", and “mAh". The corresponding images mainly feature screenshots related to system and battery settings. The third topic discusses the appearance of mobile phones, highlighting high-frequency words like “appearance", “large screen", and “color". The corresponding images primarily showcase photographs capturing the aesthetics of mobile phones. The fourth topic revolves around mobile phone photography, featuring high-frequency words such as “photography", “effect", and “camera". The corresponding images display photographic works uploaded by mobile phone users. The fifth topic pertains to mobile hardware configurations, encompassing high-frequency words such as “appearance", “large screen", and “speaker". The corresponding images include mobile screens, desktops,

and related hardware components. These five topics are crucial considerations for all brands, representing core competitive elements within the mobile phone industry.

Next, we focus on the extracted brand-specific topics, which allow us to summarize the unique characteristics, strengths, and weaknesses of each brand. We take the brands Huawei and VIVO as examples for illustration. Table 7 and Table 8 show the specific topic results for brands Huawei and VIVO, from which we can summarize the unique characteristics of these two brands.

Huawei: Focusing on System Usage Experience. Among the four specific brand topics in the Huawei corpus, three pertain to system configurations, with high-frequency terms such as “processor", “interface", and “configuration". Users of this brand are particularly concerned with the mobile phone system’s usage experience, frequently discussing the operational “speed" and the smoothness of games like “King of Glory".

VIVO: Emphasis on Appearance and Exceptional Photography Features. VIVO phones are lauded for their excellent photography capabilities. One specific topic is closely related to photography, with users focusing on functions like “blurring", “focus", and “aperture", and using the camera to capture “scene". The images associated with this topic include exquisite photos uploaded by users, such as city nightscapes and beach sunsets. VIVO users also value the phone’s appearance, with one topic highlighting “appearance", “screen", and “design".

Based on these findings, MM-LDA can also identify potential competitors for brands. For instance, in Huawei’s specific brand topics, “Apple" appears as a high-frequency word. As a brand similarly focused on high-end technological innovation, users often compare Huawei to Apple, offering valuable insights for product upgrades and improvements. Furthermore, MM-LDA can identify potential customer

Table 7 Results of Specific Topics for Brand Huawei in the ZOL Data

Specific Topic 1	Telephone,Huawei,Run,Smoothly,Fingerprint,Function,System,Feel,Speed,Software
	
Specific Topic 2	Use,Photo,Headphone,Performance,Body,Camera,Light,Battery,Design,Whole
	
Specific Topic 3	Huawei,Look,Battery,Games,Glory,System,Memory,Market,Honor of Kings,Feel
	
Specific Topic 4	Support,EMUI,Series,Inch,Processor,Interface,User,Mode,Experience,Configuration
	

Table 8 Results of Specific Topics for Brand VIVO in the ZOL Dataset

Specific Topic 1	Quality,Function,Beauty,Camera,Aperture,Front,Unlock,Fingerprint,Flash,Promotion
	   
Specific Topic 2	Effect,Experience,Blurring,Photos,Focus,Sensors,Images,Light,Camera,Scene
	   
Specific Topic 3	Telephone,Vivo,Appearance,Screen,Smooth,Feel,Like,Design,Effect,Feeling
	   

segments to enhance user retention. For example, VIVO's widely recognized selfie and beauty features reveal a potential user group focused on photography and personal appearance. By catering to the needs of this group when developing new features and upgrading products, brands can significantly boost user loyalty.

5 Experiments on Baidu Tieba

5.1 Data description

To evaluate the performance of MM-DTM, we use another dataset from Baidu Tieba, an independent brand under Baidu, which stands as a highly popular Chinese online community platform. It is built upon search engine keywords, allowing users to precisely gather around both broad and niche topics, display their flair, connect with like-minded individuals, and construct a unique platform for "interest-based" interaction. Spanning various fields such as society, regions, lifestyle, education, celebrity entertainment, gaming, sports, and business, Baidu Tieba is a leading Chinese communication platform where users freely exchange ideas.

The dataset originates from Baidu Tieba's "Valorant Bar" and "Overwatch Bar", focusing on these popular first-person shooter games with substantial player bases. Valorant, developed by Riot Games, entered its official public beta phase in 2020, while Overwatch was developed by Blizzard Entertainment and released in 2016. These two games have spurred enthusiastic discussions among netizens on Baidu Tieba. This study collected data from two periods: August 2023 (denoted as t_1) and February 2024 (denoted as t_2), employing the same methodology of exclusively har-

vesting content posted by the original posters. The total number of documents crawled by Valorant bar is 1014, and the number of documents crawled by Overwatch bar is 1099. After data collection, we built the MM-DTM to study the temporal evolution of topics and to compare how different games emphasize or shift topics over time. To process images from the dataset, we follow the procedure outlined in Section 3.2.1, converting them into collections of visual words. These processed visual words are then combined with textual words to form a unified corpus. After reprocessing, we follow the procedure used in the ZOL dataset to select the optimal number of general topics and brand-specific topics. The results show that the optimal number of general topics is 2, and each game (brand) is associated with 2 brand-specific topics.

5.2 Results of experimental mertrics

We conduct quantitative evaluations of MM-DTM by comparing it with several baseline models. Specifically, we consider four baselines: the original LDA and DTM using only textual information, a multimodal LDA model that incorporates image features into the conventional LDA framework (denoted as multi-LDA), and a multimodal DTM model that extends the traditional DTM with visual information (denoted as multi-DTM). The numbers of topics in LDA and DTM are set to match the number of general topics used in MM-DTM. To assess the quality of topics over time, we compute topic coherence and perplexity, which are commonly used evaluation metrics in topic modeling. The results are presented in Table 9. In addition, we perform a more fine-grained temporal slicing of the Baidu Tieba dataset. To this end, we implement temporal segmentation for DTM, multimodal DTM, and MM-DTM, where the original time intervals August 2023 (t_1) and February 2024 (t_2) are each equally subdivided into two finer segments (t_{11}, t_{12} and t_{21}, t_{22}). Then we rerun all the models. The results are shown in the *4-split* panel of Table 9.

Based on the results, we can draw the following conclusions. First, incorporating image information clearly improves model performance, as both LDA and DTM using only text achieve the lowest results. Second, among the multimodal models that integrate image features, MM-DTM achieves topic coherence slightly lower than other multimodal approaches (e.g., multi-LDA and multi-DTM). However, MM-DTM shows significantly better performance in terms of perplexity. This indicates that MM-DTM is particularly effective at capturing the underlying probabilistic structure of the data, and combining textual and visual information enhances both interpretability and predictive quality.

Table 9 Results of Coherence and Perplexity of Different Models on the Tieba Dataset

	Model	Coherence	Perplexity
	LDA	−3.341	1639.631
	multi-LDA	−2.048	2317.685
2 splits	DTM	−3.406	2723.226
	multi-DTM	−2.034	2539.073
	MM-DTM	−2.170	1569.033
4 splits	DTM	−2.964	1947.575
	multi-DTM	−1.777	4131.031
	MM-DTM	−2.018	1462.110

5.3 Results of general topics by MM-DTM

The general topics (shown in Table 10) reveal similarities between the two games at two time points (August 2023 and February 2024). At the first time point, topic 1 encompasses discussions around the ranked mode common to both games, reflecting the presence of a ranking system and tiers indicative of a player's skill level. Consequently, players generally aim to ascend these ranks, placing significant importance on the ranked mode. Terms like "rank", "silver", "teammate", "mentality", and "master" are central to these discussions. The corresponding images often depict snapshots of game victories or defeats, capturing moments of joy from wins or seeking consolation after losses. Topic 2 revolves around discussions related to in-game content such as skins and heroes, with terms like "new", "skin", "hero", "ability", and "buy" featuring prominently. These discussions also touch on purchasing decisions,

Table 10 The Results of General Topics of Two PC Games in Two Time Points

Time 1	Topic 1	Play, game, no, anyone, ban, really, rank, feel, good, teammates, situation, player, friends, silver, mentality, seek help, master, think, play, start, spray
		
	Topic 2	Game, one, play, hero, player, think, two, really, buy, bro, national server, match, skin, anyone, new, card, open, belt, ability
		
Time 2	Topic 1	Play, game, feel, think, rank, anyone, really, silver, diamond, gold, teammates, win, match, cheat, rank, lose, friend, season, number, drop, smurf, really, owner, block, disgusting
		
	Topic 2	One, play, no, skin, good, really, feel, season, think, pull, player, situation, Chinese server, like, have, eat, friend, match, two, buy, new
		

with players seeking community advice on Baidu Tieba, accompanied by illustrative images of skins and heroes.

At the second time point, the general topics show little change from the previous point. Topic 1 continues to focus on rank mode and gaming tiers, incorporating terms like “ban", “cheat", “lose", “disgusting", and “smurf" (meaning bullying beginners) – all reasons contributing to game losses. This fosters a communal space on Baidu Tieba for players to vent about in-game frustrations, with images reflecting match outcomes facilitating discussions based on personal gaming experiences. Topic 2 remains consistent with previous discussions about in-game skins, with images largely consisting of skins and heroes showcased by players, discussing aspects like the tactile feel of skins and their pricing.

Next, we focus on the discussed proportions of each general topic, i.e., the estimated θs. We calculate the average topic proportion for each general topic in each game dataset in the two periods. As shown in Fig. 6, for Topic 1, representing “Rank Frustration”, Valorant players exhibit greater attention in t_1 (0.5141) when compared with Overwatch players (0.2385). However, in t_2, Overwatch players’ focus on this topic increases to 0.5964, surpassing Valorant’s 0.5759. This suggests that Valorant players are more concerned about ranking frustration during the earlier period, whereas Overwatch players become more attentive in the later period, which is possibly due to game balance adjustments or other in-game factors. Similarly, for Topic 2, Overwatch players show considerably higher interest in t_1 (0.7615) when compared to Valorant players (0.4859). In contrast, in t_2, Valorant players’ interest in skins rises to 0.4241, exceeding Overwatch’s 0.4036. This shift indicates that Valorant players’ attention to skins increases in the later period, which is likely driven by the release of new cosmetics or in-game events that enhance player engagement over time.

To gain a finer-grained view of topic dynamics, we perform the same analysis on the dataset divided into 4 splits. As shown in Fig. 7, the discussion proportions of the two topics across the two games remain broadly consistent with the earlier results. We also visualize the daily evolution of the occurrence probabilities of several key words under each general topic, which can be found in Appendix G. These results reveal the temporal dynamics of word usage within topics, providing further insights into how player discussions evolve over time.

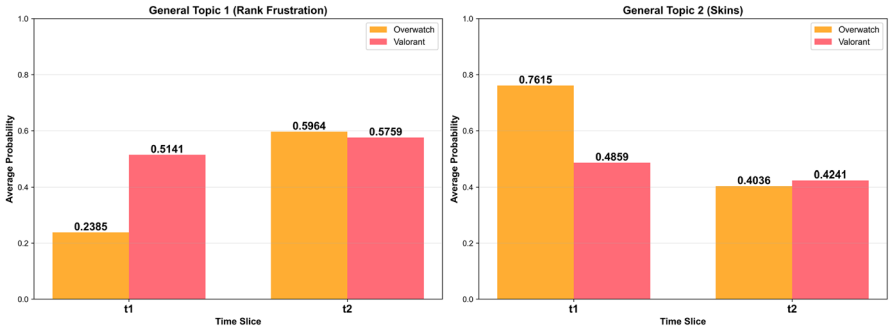


Fig. 6 The Discussed Proportions of Two General Topics in Tieba Dataset with 2 Splits

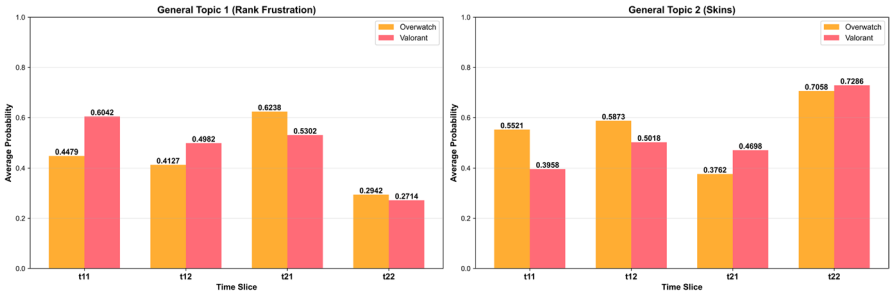


Fig. 7 The Discussed Proportions of Two General Topics in Tieba Dataset with 4 Splits

5.4 Results of specific topics by MM-DTM

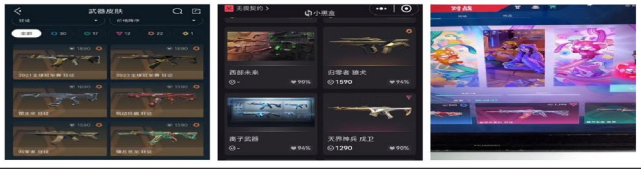
Below, we analyze the specific topics generated by MM-DTM to investigate the differences between two PC games. We first focus on the specific topics of Valorant, whose results are shown in Table 11.

At the first time point, Topic 1's vocabulary is concentrated around gaming ranks and competitive modes, with terms like "silver" and "gold" representing different gaming levels, while "record", "pressure", "win", "lose", and "ban" reflect aspects of the competitive mode. Given that ranks signify a player's gaming skill level, most players place significant emphasis on competitive mode, focusing on their records, experiencing pressure, and valuing wins and losses highly. The images under Topic 1 predominantly feature players sharing their gaming records along with discussions and rants about them. Topic 2's vocabulary captures gameplay strategies and tactics, with terms like "ice wall", "smoke", "guard", "sage" and "neon" referring to in-game characters and abilities. The corresponding images depict strategic locations for deploying character skills, with players discussing game strategies, exchanging gameplay tips, and sharing insights on Tieba to collectively improve their gaming ranks.

At the second time point, Topic 1 has not changed significantly from the previous time point, continuing to focus on gaming ranks and competitive modes. This continuity indicates an ongoing dialogue about ranks and a desire among players to team up with skilled players, as evidenced by terms like "master", "carry" and "teammate". The images largely consist of screenshots showing game outcomes, reflecting the persistent pursuit of rank improvement over time. Topic 2 shifts to discuss in-game skins, with "special effects", "skins", "night market", "pass", "champion", and "set" describing various in-game cosmetic items. Terms such as "buy", "anyone" and "help" reveal some players' hesitation between purchasing and not purchasing skins, leading them to seek opinions on Tieba. The images for this topic include screenshots of skins and the in-game store, soliciting community feedback.

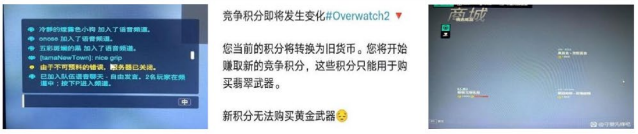
We then focus on the specific topics of Overwatch, whose results are shown in Table 12. At the first time point, a significant backdrop for the game Overwatch was its delisting from China's national servers in January 2023 and its launch on the Steam platform in August 2023, with this data point coinciding with August 2023. Topic 1 focuses on the channels through which the game is accessed, such as Battle.

Table 11 The Results of Brand-Specific Topics of Valorant in Two Time Points

Time 1	Topic 1	One, teammate, no, number, diamond, silver, kill, gold, think, block, rank, platinum, feel, win, record, pressure, in the end, bronze, ak, points, loss, true, 6, play
		
	Topic 2	Point, opposite, a, point, b, position, ice wall, contract, release, play, fearless, skill, smoke, one, guard, start, valorant, place, attack, sage, game, nanny, like, stand, fire, sand eagle, box, throw, duel, fun, neon, start, defense
		
Time 2	Topic 1	Buy, account, extraordinary, platinum, brush, diamond, play, one, bro, gold, contract, think, carry, teammate, master, god in god, want, black iron, recommendation, Hong Kong server, nanny, come on, fried fish, mythology, China server
		
	Topic 2	Fearless, contract, game, special effects, anyone, skin, buy, solve, set, tile, night market, AK, pass, gun, seek help, champion, open, free, four, segment, hour, why, novice
		

net, Steam, account, binding, accelerators, etc., all related to issues encountered during game login and their solutions. There was extensive discussion among players on how to smoothly log into the game, with experienced players sharing their tips. The images under this topic often show the errors encountered during login and issues within the game, posted on Tieba for advice on solutions. Topic 2 reflects another aspect of the situation: Overwatch's listing on Steam during this period but the previous delisting from the national servers in January, leading to the cancellation of domestic players' accounts. Players expressed their dissatisfaction by giving the game negative reviews and posting comments on Steam, with terms like negative review, platform, rating, Three Kingdoms Kill (a game with the highest negative review rate on Steam), and surpassing, illustrating this trend. Overwatch briefly became the game with the highest rate of negative reviews on Steam after its launch, surpassing Three

Table 12 The Results of Brand-Specific Topics of Overwatch in Two Time Points

Time 1	Topic 1	Battle.net, account, steam, game, bind, watch, pioneer, accelerate, login, download, register, open, connect, resolve, method, click, situation, update, file, dolphin, accelerator, enter, success, international, email, failure, program, negative review
		
	Topic 2	Watch, steam, pioneer, negative review, Blizzard, Three Kingdoms Kill, steam, ow, 2, ranking, online, review, Chinese, platform, time, gadget, mall, Battle.Net, seems to surpass
		
Time 2	Topic 1	Play, competition, season, point, steam, watch, pioneer, new, Battle.net, help, rank, server, accelerator, update, node, master, nickname, login, Blizzard, half, settlement, change
		
	Topic 2	Win, support, mechanism, environment, shadow, monster, thrill, career, encounter, moi, diamond, platinum, iron fist, hero, angel, version, new, vanguard, apocalypse, death, damage, Anna, skills, abilities, modifications, tank, fast
		

Kingdoms Kill. The images for this topic are predominantly screenshots of negative reviews, with players expressing their dissatisfaction through mockery or insults.

At the second time point, Topic 1 remains similar to the previous time, still centered on logging into the game, with terms such as Steam, Battle.net, servers, accelerators, and nodes discussing how to successfully log in. A slight difference is the mention of season updates, with terms like update, season, settlement, and new, discussing the transition from the old season to the new. The corresponding images reflect issues like network lag and server disconnections experienced after logging in, and some changes in the new season. Topic 2 focuses on discussions about the game's content and gameplay, with numerous game-related terms such as support, angel, moi, Anna, etc., representing in-game hero names and roles. From the images, which include hero changes and hero screenshots, it's evident that there were extensive discussions on hero version changes and gameplay strategies.

Based on the above results of specific topics, we find there are clear differences in the thematic content discussed for the two games. First, discussions within the Valorant community have a heavier focus on in-game skin sales. Players showcase the items available in their game store, which refreshes on Tieba, seeking opinions from the community, and some players provide reviews of the skins, offering advice to those still deciding whether to purchase. Additionally, there's more content related to gameplay strategies in Valorant, as the game was newly available in China in August 2023. This influx of new players attracted experienced players to post guides and gameplay experiences to help newcomers quickly acclimate. Sharing strategies for offense and defense, along with how to best use game items, became a hot topic among players. Second, the most notable aspect of Overwatch at this time was the surge of negative reviews on Steam upon its launch, rapidly elevating it to the top of the negative review rankings–a shocking event in gaming history. This incident became a popular topic among players, highlighting their collective frustration. However, as time progressed, some players began to move past their initial anger, focusing instead on discussing methods for logging into the game and updates to the game version, aiming to enjoy the entertainment the game offers.

6 Conclusion and discussion

In this work, we focus on brand competitive analysis using multimodal topic models. The selection of multimodal topic models is motivated by the inherent characteristics of online reviews in E-commerce platforms, where consumer opinions are increasingly expressed through a combination of textual narratives and visual evidence. Textual reviews provide explicit evaluations and rationales, while images convey complementary information such as product appearance, quality cues, and usage contexts that are often implicit or omitted in text. To jointly model these two modalities, we propose two models, i.e., MM-LDA and MM-DTM. MM-LDA enables static comparison across brands by integrating textual and visual reviews to uncover both shared and brand-specific topics. MM-DTM extends this by modeling topic evolution over time, capturing temporal trends in brand perception. These two models provide practical insights for product positioning, marketing strategy, and consumer preference tracking.

Compared with recent approaches in the literature, including unimodal topic models and neural or embedding-based multimodal methods, the proposed framework emphasizes interpretability and brand-level differentiation. Many recent neural topic models achieve strong predictive performance but rely on latent representations that are less transparent for managerial interpretation. In contrast, MM-LDA explicitly incorporates multiple brand-specific corpora within a probabilistic topic modeling framework, enabling the identification of both general topics shared across brands and topics that are unique to individual brands. This structure directly supports fine-grained cross-brand comparisons and aligns well with the needs of marketing and competitive analysis. Furthermore, by extending MM-LDA to the dynamic setting, MM-DTM captures how brand-related topics emerge and evolve over time. This aspect is often overlooked in existing multimodal topic modeling studies, since

the latter primarily focuses on static analyses. Together, these two models offer a more comprehensive and faithful representation of consumer perceptions than existing topic models, which is particularly important for longitudinal brand competitive analysis.

This work still has several limitations. First, our current model performance is verified based on a relatively simple visual feature extraction method. Incorporating more advanced visual representations would further enhance the robustness of our evaluation. Second, in the current study, we do not incorporate additional signals such as sentiment, label, and product sequences. However, these factors are highly relevant for improving the interpretability and robustness of multimodal topic models. Therefore, extending our framework to include this information constitutes an important direction for future research. Third, in our current study, the data for different brands are collected from the same platform, which ensures a certain level of consistency across corpora. However in practical scenarios, brand-specific data may originate from different platforms, introducing potential heterogeneity due to variations in comment organization, presentation, or platform-specific algorithms. Addressing such heterogeneity represents an important direction for future research. Finally, while the current evaluation focuses on capturing the temporal evolution of topics in multimodal reviews, future work could extend the framework to predictive modeling of emerging trends. Incorporating such predictive analyses would further demonstrate the practical utility of MM-DTM, which we plan to explore in future work.

Appendix A: Model estimation of MM-LDA

We employ the Gibbs sampling method [7] for estimation of MM-LDA. Let $\boldsymbol{\Theta}^{\boldsymbol{g}}$ and $\boldsymbol{\Theta}^{\boldsymbol{s}}$ represent the collection of document–topic distributions over general and specific topics, respectively. Let $\boldsymbol{\Phi}_{\boldsymbol{k}}^{\boldsymbol{T,g}}$, $\boldsymbol{\Phi}_{\boldsymbol{k}}^{\boldsymbol{I,g}}$, $\boldsymbol{\Phi}_{\boldsymbol{k,b}}^{\boldsymbol{T,s}}$, and $\boldsymbol{\Phi}_{\boldsymbol{k,b}}^{\boldsymbol{I,s}}$ denote the collections of word distributions of general and brand-specific topics over textual and visual vocabularies, respectively. Let $\boldsymbol{w}$ represent all observed textual words and image visual words in the corpora, and let $\boldsymbol{u}$ and $\boldsymbol{z}$ represent the corresponding indicator variables. Given the observed data $\boldsymbol{w}$ and hyperparameters $(\beta, \beta_b, \alpha_g, \alpha_s, \gamma)$, we first derive the complete posterior distribution according to the generative process of MM-LDA:

$$
\begin{aligned}
&P(\mathrm{z}^s, \mathrm{z}^g, \mathrm{u}, \boldsymbol{\theta}^s, \boldsymbol{\theta}^g, \boldsymbol{\Phi}^{T,s}_{k,b}, \boldsymbol{\Phi}^{I,s}_{k,b}, \boldsymbol{\Phi}^{T,g}_{k}, \boldsymbol{\Phi}^{I,g}_{k}, p \mid \alpha, \beta, \gamma) \propto \\
&\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{k=1}^{K_0}(\theta^g_{d,b,k})^{\alpha-1}\Big]\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{k=1}^{K_b}(\theta^s_{d,b,k})^{\alpha-1}\Big]\Big[\prod_{k=1}^{K_0}\prod_{v=1}^{V^T}(\Phi^{T,g}_{k,v})^{\beta-1}\Big]\Big[\prod_{k=1}^{K_0}\prod_{v=1}^{V^I}(\Phi^{I,g}_{k,v})^{\beta-1}\Big] \\
&\Big[\prod_{b=1}^{B}\prod_{k=1}^{K_b}\prod_{v=1}^{V^T_b}(\Phi^{T,s}_{k,b,v})^{\beta-1}\Big]\Big[\prod_{b=1}^{B}\prod_{k=1}^{K_b}\prod_{v=1}^{V^I_b}(\Phi^{I,s}_{k,b,v})^{\beta-1}\Big]\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{N_{d,b}}p_{d,b,n}^{\gamma-1}(1-p_{d,b,n})^{\gamma-1}\Big] \\
&\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{N_{d,b}}\mathrm{Beta}(\gamma)\, p_{d,b,n}^{u^T_{d,b,n}}(1-p_{d,b,n})^{1-u^T_{d,b,n}}\Big]\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{M_{d,b}}\mathrm{Beta}(\gamma)\, p_{d,b,n}^{u^I_{d,b,n}}(1-p_{d,b,n})^{1-u^I_{d,b,n}}\Big] \\
&\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{N_{d,b}}(\Phi^{T,g}_{z^g_{d,b,n},v_{d,b,n}})^{I(u^T_{d,b,n}=0)}\Big]\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{N_{d,b}}(\Phi^{T,s}_{z^s_{d,b,n},b,v_{d,b,n}})^{I(u^T_{d,b,n}=1)}\Big] \\
&\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{M_{d,b}}(\Phi^{I,g}_{z^g_{d,b,n},v_{d,b,n}})^{I(u^I_{d,b,n}=0)}\Big]\Big[\prod_{b=1}^{B}\prod_{d=1}^{D_b}\prod_{n=1}^{M_{d,b}}(\Phi^{I,s}_{z^s_{d,b,n},b,v_{d,b,n}})^{I(u^I_{d,b,n}=1)}\Big].
\end{aligned}
\tag{A.1}
$$

Based on the above posterior distribution, we obtain the following conditional distributions. Iteratively sampling from these conditional distributions until convergence yields the final estimation. Below, we present the conditional distributions for textual information. The conditional distributions for image information are similar and thus omitted. First, for the textual word *n* in document *d* of corpus *b*, given the values of all other latent variables, the conditional distribution of the general topic indicator $z^g_{d_b,n}$ is:

$$
P(z^g_{d,b,n}=k \mid \boldsymbol{z}^g_{\neg(d,b,n)}, \boldsymbol{u}, \boldsymbol{w}) \propto \frac{c^{d,b,g}_{(k)}+\alpha}{c^{d,b,g}_{(\cdot)}+K_0\alpha} \times \frac{c^{g,k}_{(v)}+\beta}{c^{g,k}_{(\cdot)}+V^T\beta}. \tag{A.2}
$$

Here, $c^{d_b,g}(k)$ is the count of textual words assigned to topic *k* in document d_b, and $c^{d_b,g}(.)$ is the total count of textual words in document d_b; $c^{g,k}(v)$ is the count of textual word *v* assigned to topic *k*, and $c^{g,k}(.)$ is the total count of any textual word assigned to topic *k*. $\boldsymbol{z}^g_{\neg(d_b,n)}$ denotes the collection of $\boldsymbol{z}^g$ excluding $z^g_{d_b,n}$. All counts denoted by *c* do not include the *n* word in document *d*. Second, the conditional distribution for the specific textual topic indicator $z^s_{d_b,s}$ and image topic indicator $z^s_{d_b,s}$ is given by the following equation. The notation is similar to the above, with the only difference being the superscript *s* indicating specific words and the superscript *b* indicating that the counts are restricted to corpus *b*.

$$
P(z^s_{d,b,n}=k \mid \boldsymbol{z}^s_{\neg(d,b,n)}, \boldsymbol{u}, \boldsymbol{w}) \propto \frac{c^{d,b,s}_{(k)}+\alpha}{c^{d,b,s}_{(\cdot)}+K_b\alpha} \times \frac{c^{s,k,b}_{(v)}+\beta}{c^{s,k,b}_{(\cdot)}+V^T_b\beta}. \tag{A.3}
$$

Third, the conditional probabilities of the indicator variables $u^T_{d_b,n}$ and $u^I_{d_b,n}$ are derived as follows,

$$P(u^{l}_{d,b,n} = x \mid \boldsymbol{z}, \boldsymbol{u}_{\neg(d,b,n)}, \boldsymbol{w}) \propto \begin{cases} \dfrac{c^{T}_{(0)} + \gamma}{c^{T}_{(\cdot)} + 2\gamma}, & \text{if } l = T,\ x = 0, \\ \dfrac{c^{I}_{(0)} + \gamma}{c^{I}_{(\cdot)} + 2\gamma}, & \text{if } l = I,\ x = 0, \\ \dfrac{c^{T}_{(1)} + \gamma}{c^{T}_{(\cdot)} + 2\gamma}, & \text{if } l = T,\ x = 1, \\ \dfrac{c^{I}_{(1)} + \gamma}{c^{I}_{(\cdot)} + 2\gamma}, & \text{if } l = I,\ x = 1. \end{cases} \tag{A.4}$$

Here $c^{T}_{(0)}$ represents the count of textual words belonging to specific topics, $c^{T}_{(1)}$ represents the count of textual words belonging to general topics, and $c^{T}_{(\cdot)}$ represents the total count of textual words. The corresponding values with the superscript *I* have similar meanings but refer to visual word counts. All counts are computed for document *d* by excluding the *n*-th word.

Appendix B: Model estimation of MM-DTM

The static part of the model uses Gibbs sampling method for parameter inference. Firstly, Gibbs sampling requires the joint density of parameters, which is calculated through integration:

$$P(\boldsymbol{W}^{T}, \boldsymbol{W}^{I}, \boldsymbol{Z}, \boldsymbol{U} \mid \Theta) = \iint P(\boldsymbol{W}^{T}, \boldsymbol{W}^{I}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{\Theta}^{g}, \boldsymbol{\Theta}^{s}, \boldsymbol{\Phi}^{T,g}, \boldsymbol{\Phi}^{I,g}, \boldsymbol{\Phi}^{T,s}, \boldsymbol{\Phi}^{I,s}, p \mid \Theta)\ d\boldsymbol{\Theta}\, dp$$

Where $\vec{Z}$ denotes topics of all textual words and visual words, $\vec{U}$ denotes indicators of all textual words and visual words, Θ denotes all hyperparameters, θ denotes the topic distribution, β and $\vec{\phi}$ denotes the topic-word distribution, and p denotes the bernoulli probability of indicators $\vec{U}$.

Then, the density functions for each part of the textual segment are articulated as follows:

$$P(\boldsymbol{W}^T \mid \boldsymbol{\Phi}) = \prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{w=1}^{N_{d_b}} P(W_{d_b,w} \mid \boldsymbol{\Phi}^{T,g}, \boldsymbol{\Phi}^{T,s})$$

$$= \left[\prod_{b=1}^{B} \prod_{k_b=1}^{K_b^s} \prod_{r=1}^{V_w} \left(\Phi_{k_b,b,r}^{T,s}\right)^{n_{(\cdot),r}^{k_b,s}} \right] \cdot \left[\prod_{k=1}^{K_0} \prod_{r=1}^{V_w} \left(\Phi_{k,r}^{T,g}\right)^{n_{(\cdot),r}^{k,g}} \right]$$

$$P(\vec{Z}_w \mid \vec{\theta}) = \prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{w=1}^{N_{d_b}} P\left(Z_{d_b,w} \mid \vec{\theta}_{d_b}\right)$$

$$= \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{k_b=1}^{K_b^s} \left(\theta_{d_b,k_b}^{s}\right)^{n_{d_b,(\cdot)}^{k_b,s}} \right] \cdot \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{k=1}^{K_0} \left(\theta_{d_b,k}^{g}\right)^{n_{d_b,(\cdot)}^{k,g}} \right]$$

$$P\left(\vec{U}_w \mid p\right) = \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{v=1}^{M_{d_b}} P(U_{d_b,v} \mid p) \right] = p^{n_{(\cdot)}^1} (1-p)^{n_{(\cdot)}^0}$$

Where the count variable $n_{3,4}^{1,2}$ represents the count of textual words, where the first position denotes the topic number, the second position *s* or *g* indicates a general or specific topic, the third position represents the document number, and the fourth position denotes the word number. The symbol $(\cdot)$ represents summation over all values at that position. Similarly, *m* denotes the count of visual words. And the variable $n_{(\cdot)}^1$ represents the summation of text word counts under all general topics, while $n_{(\cdot)}^0$ denotes the summation of text word counts under all specific topics. Similarly, $m_{(\cdot)}^1$ and $m_{(\cdot)}^0$ represent the count of visual words.

Additionally, the density functions for each part of the visual word segment are outlined as follows:

$$P(\boldsymbol{W}^I \mid \boldsymbol{\Phi}) = \left[\prod_{b=1}^{B} \prod_{k_b=1}^{K_b^s} \prod_{r=1}^{V_v} \left(\Phi_{k_b,b,r}^{I,s}\right)^{m_{(\cdot),r}^{k_b,s}} \right] \cdot \left[\prod_{k=1}^{K_0} \prod_{r=1}^{V_v} \left(\Phi_{k,r}^{I,g}\right)^{m_{(\cdot),r}^{k,g}} \right]$$

$$P(\vec{Z}_v \mid \vec{\theta}) = \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{k_b=1}^{K_b^s} \left(\theta_{d_b,k_b}^{s}\right)^{m_{d_b,(\cdot)}^{k_b,s}} \right] \cdot \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{k=1}^{K_0} \left(\theta_{d_b,k}^{g}\right)^{m_{d_b,(\cdot)}^{k,g}} \right]$$

$$P\left(\vec{U}_v \mid p\right) = \left[\prod_{b=1}^{B} \prod_{d_b=1}^{D_b} \prod_{v=1}^{M_{d_b}} P(U_{d_b,v} \mid p) \right] = p^{m_{(\cdot)}^1} (1-p)^{m_{(\cdot)}^0}$$

Where the word Indicator U_{bij}: The indicator for the j^{th} word in the i^{th} document under brand *b*, determining whether the word belongs to a general or specific topic. The subscripts *w* and *v* indicate whether it is a textual or visual word.

And the density function of the prior distribution:

$$P\left(\vec{\theta}_d^s, \vec{\theta}_d^g \mid \alpha\right) = \left[\prod_{b=1}^{B}\prod_{d_b=1}^{D_b}\frac{\Gamma(K_b^s\alpha)}{\prod_{k_b=1}^{K_b^s}\Gamma(\alpha)}\prod_{k_b=1}^{K_b^s}\left(\theta_{d_b,k_b}^s\right)^{\alpha-1}\right] \cdot \left[\prod_{b=1}^{B}\prod_{d_b=1}^{D_b}\frac{\Gamma(K_0\alpha)}{\prod_{k=1}^{K_0}\Gamma(\alpha)}\prod_{k=1}^{K_0}\left(\theta_{d_b,k}^g\right)^{\alpha-1}\right]$$

$$P(p \mid \gamma) = \frac{p^{\gamma-1}(1-p)^{\gamma-1}}{B(\gamma,\gamma)}$$

Multiplying them together and the joint probability obtained by integration is:

$$\left[\prod_{b=1}^{B}\prod_{k_b=1}^{K_b^s}\prod_{r=1}^{V_v}\left(\Phi_{k_b,b,r}^{I,s}\right)^{m_{(\cdot),r}^{k_b,s}}\right] \cdot \left[\prod_{k=1}^{K_0}\prod_{r=1}^{V_v}\left(\Phi_{k,r}^{I,g}\right)^{m_{(\cdot),r}^{k,g}}\right] \cdot \left[\prod_{b=1}^{B}\prod_{k_b=1}^{K_b^s}\prod_{r=1}^{V_w}\left(\Phi_{k_b,b,r}^{T,s}\right)^{n_{(\cdot),r}^{k_b,s}}\right]$$
$$\cdot \left[\prod_{k=1}^{K_0}\prod_{r=1}^{V_w}\left(\Phi_{k,r}^{T,g}\right)^{n_{(\cdot),r}^{k,g}}\right] \cdot \frac{B\left(\gamma + n_{(\cdot)}^1 + m_{(\cdot)}^1,\ \gamma + n_{(\cdot)}^0 + m_{(\cdot)}^0\right)}{B(\gamma,\gamma)}$$
$$\cdot \left[\prod_{b=1}^{B}\prod_{d_b=1}^{D_b}\left(\frac{\Gamma(K_b^s\alpha)}{\prod_{k_b=1}^{K_b^s}\Gamma(\alpha)}\frac{\prod_{k_b=1}^{K_b^s}\Gamma\left(n_{d_b,(\cdot)}^{k_b,s} + m_{d_b,(\cdot)}^{k_b,s} + \alpha\right)}{\Gamma\left(n_{d_b,(\cdot)}^{(\cdot),s} + m_{d_b,(\cdot)}^{(\cdot),s} + K_b^s\alpha\right)}\right)\right]$$
$$\cdot \left[\prod_{b=1}^{B}\prod_{d_b=1}^{D_b}\left(\frac{\Gamma(K_0\alpha)}{\prod_{k=1}^{K_0}\Gamma(\alpha)}\frac{\prod_{k=1}^{K_0}\Gamma\left(n_{d_b,(\cdot)}^{k,g} + m_{d_b,(\cdot)}^{k,g} + \alpha\right)}{\Gamma\left(n_{d_b,(\cdot)}^{(\cdot),g} + m_{d_b,(\cdot)}^{(\cdot),g} + K_0\alpha\right)}\right)\right]$$

In the above formula, *n* and *m* respectively represent the counts of textual words and visual words.

We define variational parameters for topic distributions $r_{bijk,g}^w, r_{bijk,s}^w, r_{bijk,g}^v, r_{bijk,s}^v$. The variational probability that the j^{th} word in the i^{th} document under brand b belongs to the k^{th} topic, with superscripts *w* and *v* indicating textual or visual words, respectively. The *s* and *g* denote whether it is a specific or general topic.

Following the Collapsed Gibbs algorithm for deriving the update formula, the estimates for $\mathrm{r_{bijk,s}}$ and $\mathrm{r_{bijk,g}}$ can be obtained. These estimates are directly substituted into the joint probability. The counts below are excluding the count of the word being updated, assuming the count of this word does not exist:

$$r_{bijk,g}^w = P\left(Z_{bij,g}^w = k\right) = P\left(Z_{bij,g}^w = k \mid \vec{U}, \vec{W}, \vec{Z}\neg bij, g, \Theta\right)$$
$$= I(U_{bij,w} = 1) \cdot \frac{\alpha + n_{i,(\cdot)}^{k,g} + m_{i,(\cdot)}^{k,g}}{K_0\alpha + n_{i,(\cdot)}^{(\cdot),g} + m_{i,(\cdot)}^{(\cdot),g}} \cdot \Phi_{k,w}^{T,g}$$

Similarly, other update formulas can be derived as follows:

$$r^{w}_{bijk,s} = I(U_{bij,w} = 0) \cdot \frac{\alpha + n^{k,s}_{i,(\cdot)} + m^{k,s}_{i,(\cdot)}}{K^{s}_{b}\alpha + n^{(\cdot),s}_{i,(\cdot)} + m^{(\cdot),s}_{i,(\cdot)}} \cdot \Phi^{T,s}_{k_b,w}$$

$$r^{v}_{bijk,g} = I(U_{bij,v} = 1) \cdot \frac{\alpha + n^{k,g}_{i,(\cdot)} + m^{k,g}_{i,(\cdot)}}{K_{0}\alpha + n^{(\cdot),g}_{i,(\cdot)} + m^{(\cdot),g}_{i,(\cdot)}} \cdot \Phi^{I,g}_{k,v}$$

$$r^{v}_{bijk,s} = I(U_{bij,v} = 0) \cdot \frac{\alpha + n^{k,s}_{i,(\cdot)} + m^{k,s}_{i,(\cdot)}}{K^{s}_{b}\alpha + n^{(\cdot),s}_{i,(\cdot)} + m^{(\cdot),s}_{i,(\cdot)}} \cdot \Phi^{I,s}_{k_b,v}$$

$$P(U_{bij,w} = 1) = \frac{\alpha + n^{k,g}_{i,(\cdot)} + m^{k,g}_{i,(\cdot)}}{K_{0}\alpha + n^{(\cdot),g}_{i,(\cdot)} + m^{(\cdot),g}_{i,(\cdot)}} \cdot \frac{\gamma + n^{1}_{(\cdot)} + m^{1}_{(\cdot)}}{2\gamma + n_{(\cdot)} + m_{(\cdot)}} \cdot \Phi^{T,g}_{k,w}$$

$$P(U_{bij,w} = 0) = \frac{\alpha + n^{k,s}_{i,(\cdot)} + m^{k,s}_{i,(\cdot)}}{K^{s}_{b}\alpha + n^{(\cdot),s}_{i,(\cdot)} + m^{(\cdot),s}_{i,(\cdot)}} \cdot \frac{\gamma + n^{0}_{(\cdot)} + m^{0}_{(\cdot)}}{2\gamma + n_{(\cdot)} + m_{(\cdot)}} \cdot \Phi^{T,s}_{k_b,w}$$

$$P(U_{bij,v} = 1) = \frac{\alpha + n^{k,g}_{i,(\cdot)} + m^{k,g}_{i,(\cdot)}}{K_{0}\alpha + n^{(\cdot),g}_{i,(\cdot)} + m^{(\cdot),g}_{i,(\cdot)}} \cdot \frac{\gamma + m^{1}_{(\cdot)} + n^{1}_{(\cdot)}}{2\gamma + n_{(\cdot)} + m_{(\cdot)}} \cdot \Phi^{I,g}_{k,v}$$

$$P(U_{bij,v} = 0) = \frac{\alpha + n^{k,s}_{i,(\cdot)} + m^{k,s}_{i,(\cdot)}}{K^{s}_{b}\alpha + n^{(\cdot),s}_{i,(\cdot)} + m^{(\cdot),s}_{i,(\cdot)}} \cdot \frac{\gamma + n^{0}_{(\cdot)} + m^{0}_{(\cdot)}}{2\gamma + n_{(\cdot)} + m_{(\cdot)}} \cdot \Phi^{I,s}_{k_b,v}$$

In the update formula for U, where k represents the corresponding topic for the word, the estimation formula for the word distribution can subsequently be derived as follows:

$$\Phi^{T,s}_{k,b,v} \propto \frac{n^{k,s}_{(\cdot),v}}{n^{k,s}_{(\cdot),(\cdot)}}, \quad \Phi^{T,g}_{k,v} \propto \frac{n^{k,g}_{(\cdot),v}}{n^{k,g}_{(\cdot),(\cdot)}}, \quad \Phi^{I,s}_{k,b,v} \propto \frac{m^{k,s}_{(\cdot),v}}{m^{k,s}_{(\cdot),(\cdot)}}, \quad \Phi^{I,g}_{k,v} \propto \frac{m^{k,g}_{(\cdot),v}}{m^{k,g}_{(\cdot),(\cdot)}}.$$

For the second part, the dynamic aspect of model update formulas inference: Assuming $\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}\big(\beta_{t-1,k},\, \sigma^2 I\big)$, a variational distribution for β is then introduced:

$$\widehat{\beta}_{t,k} \mid \beta_{t,k} \sim \mathcal{N}\big(\beta_{t,k},\, \hat{v}_t^2 I\big)$$

Below, the subscript k for the topic is omitted because the same process is applied to each word distribution β, and different word distributions β are independent of each other. The overall likelihood only needs a product from $k = 1$ to K, and then the last update of β by taking the partial derivative will eliminate the product, leaving only the current topic k's term. Therefore, only a specific topic is considered below.

We use $\widetilde{m}_{tw}$ and $\widetilde{V}_{tw}$ to denote the backward mean and the backward variance of Kalman filtering algorithm, while m_{tw} and V_{tw} denote the forward mean and the forward variance. And $\widehat{\beta}_{tw}$ denotes the variational distribution of β_{tw} and $\hat{v}_t^2$ denotes the

variance of the variational distribution. Besides, σ^2 denotes the variance of the evolution of word distribution which characterized as a conditional normal distribution.

This part is the same as the parameter inference of the DTM model, using conjugate gradient method to update the parameter:

$$\frac{\partial l(\widehat{\beta},\hat{v})}{\partial \widehat{\beta}_{tw}} = -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left(\widetilde{m}_{tw}-\widetilde{m}_{t-1,w}\right)\left(\frac{\partial \widetilde{m}_{tw}}{\partial \widehat{\beta}_{tw}}-\frac{\partial \widetilde{m}_{t-1,w}}{\partial \widehat{\beta}_{tw}}\right) + \sum_{t=1}^{T}\left(n_{tw}-n_t\hat{\zeta}_t^{-1}\exp\left(\widetilde{m}_{tw}+\frac{\widetilde{V}_{tw}}{2}\right)\right)\frac{\partial \widetilde{m}_{tw}}{\partial \widehat{\beta}_{tw}}$$

Where $n_t=\sum_w n_{tw},\ \hat{\zeta}_t=\sum_w \exp\left(\widetilde{m}_{tw}+\frac{\widetilde{V}_{tw}}{2}\right)$.

Given $\partial m_0/\partial\widehat{\beta}_s=0$, the forward derivative is calculated as:

$$\frac{\partial m_t}{\partial \widehat{\beta}_s}=\left(\frac{\hat{v}_t^2}{\hat{v}_t^2+\sigma^2+V_{t-1}}\right)\frac{\partial m_{t-1}}{\partial \widehat{\beta}_s}+\left(1-\frac{\hat{v}_t^2}{\hat{v}_t^2+\sigma^2+V_{t-1}}\right)\delta_{s,t}$$

Where $\delta_{s,t}$ is an indicator function, taking value 1 when $s=t$.

Given $\partial\widetilde{m}_T/\partial\widehat{\beta}_s=\partial m_T/\partial\widehat{\beta}_s$, the backward derivative is calculated as:

$$\frac{\partial \widetilde{m}_{t-1}}{\partial \widehat{\beta}_s}=\left(\frac{\sigma^2}{\sigma^2+V_{t-1}}\right)\frac{\partial m_{t-1}}{\partial \widehat{\beta}_s}+\left(1-\frac{\sigma^2}{\sigma^2+V_{t-1}}\right)\frac{\partial \widetilde{m}_t}{\partial \widehat{\beta}_s}$$

Appendix C: The detailed setting of the detailed setting of candidate models in ZOL data analysis

Table 13 The Number of Topics in Candidate Models in ZOL Data

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
General Topics	1	2	2	3	3	3	3	4
Huawei	1	1	2	2	2	2	2	2
VIVO	1	1	1	1	1	1	2	2
Samsung	1	1	1	1	1	2	2	2
OPPO	1	1	1	1	2	2	2	2
	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	
General Topics	4	5	5	5	5	5	6	
Huawei	2	2	3	3	4	4	4	
VIVO	2	2	2	2	2	3	3	
Samsung	3	3	3	4	4	4	4	
OPPO	2	2	2	2	2	2	2	

Appendix D: Using the CLIP method

The CLIP (Contrastive Language–Image Pretraining) model is a large-scale vision–language representation framework that jointly learns textual and visual embeddings through contrastive learning on hundreds of millions of image–text pairs [4, 13]. It aligns image and text features in a shared semantic space, enabling rich multimodal understanding and cross-modal retrieval. To assess whether such deep visual representations improve the performance of MM-LDA, we replace the SIFT-based image features with the CLIP embeddings. The results show that, the CLIP-based variant achieves much lower topic coherence (-2.246 vs. -1.453) and comparable perplexity (1058.192 vs. 1180.975) than MM-LDA. This indicates that while CLIP provides more abstract and semantically aligned features, traditional handcrafted descriptors such as SIFT may better capture concrete visual distinctions relevant to product-level topics. Overall, the results confirm that our MM-LDA framework remains robust under different visual feature representations.

Appendix E: Sensitivity analysis

To evaluate the robustness of our model, we first conduct sensitivity analysis by varying the hyperparameters α, β_{image}, β_{text}, and γ within the range [0.01, 1.0], while keeping other settings fixed. As shown in Table 14, both coherence and perplexity remain relatively stable across different parameter values, indicating that the model is not highly sensitive to hyperparameter fluctuations. Minor variations suggest that smaller β values may slightly reduce topic quality, whereas excessively large values lead to more homogeneous topics. The chosen configuration ($\alpha = 0.1$, $\beta_{\text{image}} = 0.01$, $\beta_{\text{text}} = 0.01$, $\gamma = 0.3$) achieves an optimal balance between coherence and perplexity, confirming its suitability for the main experiments.

Table 14 Effects of Hyperparameters on Topic Coherence and Perplexity

Metric	α		β_{image}		β_{text}		γ	
	Value	Score	Value	Score	Value	Score	Value	Score
Coherence	0.01	−1.852	0.001	−1.959	0.001	−1.888	0.1	−1.824
	0.1	−1.966	0.01	−1.823	0.01	−1.840	0.3	−1.937
	0.5	−1.954	0.5	−1.941	0.5	−1.537	0.7	−1.886
	1.0	−1.845	1.0	−1.997	1.0	−1.550	0.9	−1.918
Perplexity	0.01	1506.641	0.001	1409.386	0.001	1444.480	0.1	1318.120
	0.1	1335.957	0.01	1262.033	0.01	1202.283	0.3	1463.338
	0.5	1204.639	0.5	1325.133	0.5	1339.371	0.7	1258.474
	1.0	1380.430	1.0	1636.889	1.0	1445.502	0.9	1339.935

Table 15 Summary Statistics of Coherence and Perplexity in 20 Replications for MM-LDA

Metric	Min	Mean	SD	Max
Coherence	−1.630	−1.462	0.038	−1.389
Perplexity	1140.838	1280.739	41.739	1391.557

Table 16 Results of Coherence and Perplexity of Different Models on the Static Setting of Baidu Tieba Dataset

Metric	LDA	BJ-LDA	Multimodal-ZeroShotTM	MM-LDA
Coherence	−3.644	−3.423	−13.233	−2.352
Perplexity	1870.765	2058.546	2081.057	1240.268

Appendix F: Stability check

To assess the stability of MM-LDA, we conduct 20 independent replications using the same experimental setup. The results are summarized in Table 15. We find that both coherence and perplexity remain highly consistent across runs, with very small standard deviations. These results indicate that the MM-LDA model can produce stable and reproducible results.

To further validate the performance of our approach on a larger-scale corpus, we apply MM-LDA and several baseline models (e.g., LDA, BJ-LDA, and Multimodal-ZeroShotTM) to the Baidu Tieba dataset, which is treated as a static collection. As shown in Table 16, our proposed MM-LDA continues to outperform the baselines in terms of both topic coherence and perplexity, confirming its robustness and generalizability to larger corpora.

Appendix G: The average frequency trend of topic keywords

See Figs. 8 9 10 11

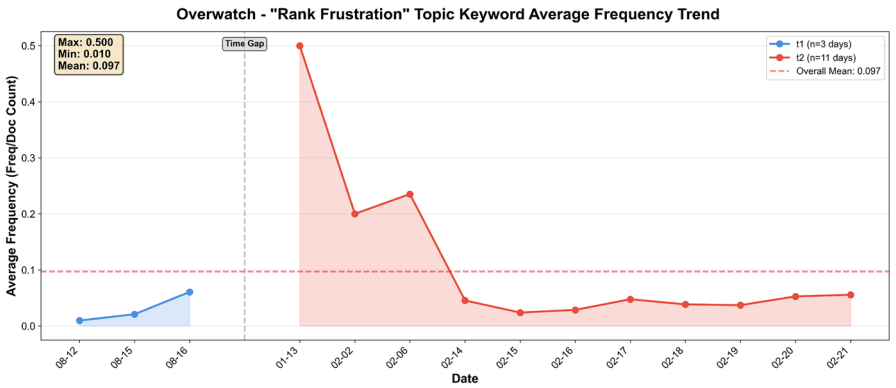


Fig. 8 The Daily Frequency Trend of Keyword “Rank Frustration" in General Topic 1 for Overwatch

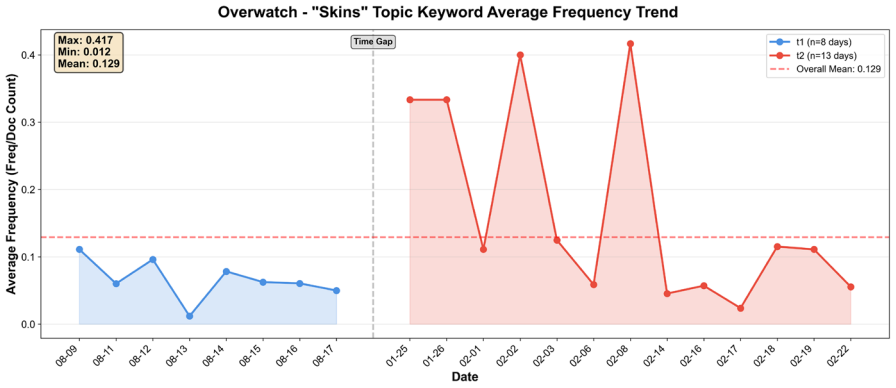


Fig. 9 The Daily Frequency Trend of Keyword “Skins" in General Topic 2 for Overwatch

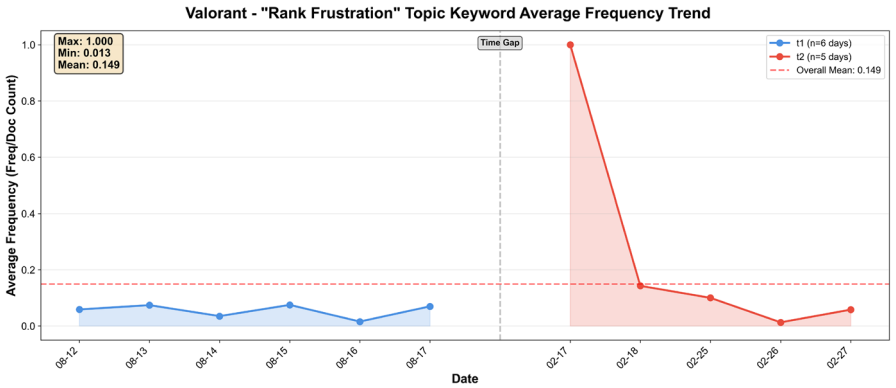


Fig. 10 The Daily Frequency Trend of Keyword “Rank Frustration" in General Topic 1 for Valorant

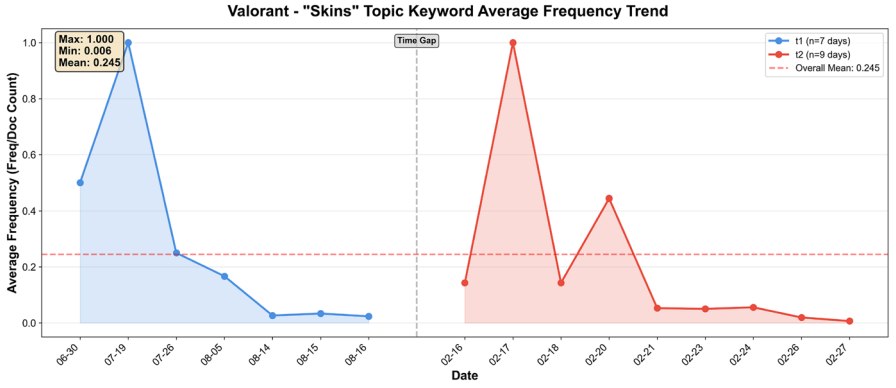


Fig. 11 The Daily Frequency Trend of Keyword “Skins" in General Topic 2 for Valorant

Acknowledgements This research is supported by Fundamental Research Funds for the Central Universities (CUC25GT23), National Natural Science Foundation of China (72371241, 72171229), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China, and Public Computing Cloud, Renmin University of China.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest to disclose.

References

1. Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,3*, 993–1022. 18th International Conference on Machine Learning, WILLIAMSTOWN, MA, JUN 28-JUL 01, 2001.
2. Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *ACM Sigir Forum* (pp. 127–134).
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research,3*, 993–1022.
4. Gonzalez-Pizarro, F., & Carenini, G. (2024). Neural multimodal topic modeling: A comprehensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 12159–12172). Turin (Torino), Italy: ELRA & ICCL anthology ID: 2024.lrec-main.1064.
5. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America,101*, 5228–5235.
6. Guo, Y., Wang, F., Xing, C., & Lu, X. (2022). Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent Dirichlet allocation. *Electronic Commerce Research and Applications,53*, 101141.
7. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.
8. Huang, X., Wan, X., & Xiao, J. (2011). Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 648–653).
9. Kato, H., & Harada, T. (2014). Image reconstruction from bag-of-visual-words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 955–962).
10. Li, C., Kwok, L., Xie, K. L., Liu, J., & Ye, Q. (2023). Let photos speak: The effect of user-generated visual content on hotel review helpfulness. *Journal of Hospitality & Tourism Research,47*, 665–690.
11. Li, H., Ji, H., Liu, H., Cai, D., & Gao, H. (2022). Is a picture worth a thousand words? Understanding the role of review photo sentiment and text-photo sentiment disparity using deep learning algorithms. *Tourism Management,92*, 104559.
12. Li, H., Qian, Y., Jiang, Y., Liu, Y., & Zhou, F. (2023). A novel label-based multimodal topic model for social media analysis. *Decision Support Systems,164*, 113863.
13. Liu, J., Zhang, L., & Yan, Q. (2024). I-Topic: An Image-text Topic Modeling Method Based on Community Detection. In *2024 5th International Conference on Computer Engineering and Application (ICCEA)* (pp. 797–800).
14. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,60*, 91–110.
15. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach,, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32 (NIPS 2019) vol. 32 of Advances in Neural Information Processing Systems, 33rd Conference on Neural Information Processing Systems (NeurIPS)* CANADA: Vancouver, DEC 08-14, 2019.
16. Lu, X., Wang, F., Zhou, R., & Feng, Y. (2022). Bayesian sparse joint dynamic topic model with flexible lead-lag order. *Information Sciences,614*, 392–410.

17. Luo, C., Luo, X. R., Xu, Y., Warkentin, M., & Sia, C. L. (2015). Examining the moderating role of sense of membership in online review evaluations. *Information & Management,52*, 305–316.
18. Ma, Y., Xiang, Z., Du, Q., & Fan, W. (2018). Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *International Journal of Hospitality Management,71*, 120–131.
19. Putthividhya, D., Attias, H. T., & Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Comp Soc, IEEE Conference on Computer Vision and Pattern Recognition, 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3408–3415). San Francisco, CA, JUN 13-18, 2010.
20. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila, & T. Zhang (Eds.), *International Conference on Machine Learning, VOL 139, vol. 139 of Proceedings of Machine Learning Research, International Conference on Machine Learning (ICML)* ELECTR NETWORK, JUL 18-24, 2021.
21. Shao, X., Tang, G., & Bao, B.-K. (2019). Personalized travel recommendation based on sentiment-aware multimodal topic model. *IEEE Access,7*, 113043–113052.
22. Sipos, R., & Joachims, T. (2013). Generating comparative summaries from reviews. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 1853–1856).
23. Villarroel Ordenes, F., & Zhang, S. (2019). From words to pixels: Text and image mining methods for service research. *Journal of Service Management,30*, 593–620.
24. Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *Uai*, 579–586.
25. Wang, D., Zhu, S., Li, T., & Gong, Y. (2013). Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data (TKDD),7*, 1–18.
26. Wang, F., Yuan, X., & Lu, X. (2025). Knowledge emerging trend detection using relevance-based dynamic thin topic model. *Pattern Analysis and Applications,28*, 116.
27. Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications,29*, 142–156.
28. Xia, X., Zhao, Y., & Jiang, D. (2022). Multimodal interaction enhanced representation learning for video emotion recognition. *Frontiers in Neuroscience,16*, 1086380.
29. Xu, N., Mao, W., & Chen, G. (2019). Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence,33*, 371–378.
30. Xue, F., Hong, R., He, X., Wang, J., Qian, S., & Xu, C. (2019). Knowledge-based topic model for multi-modal social event analysis. *IEEE Transactions on Multimedia,22*, 2098–2110.
31. Yang, J., Jiang, Y.-G., Hauptmann, A. G., & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval* (pp. 197–206).
32. You, Q., Luo, J., Jin, H., & Yang, J. (2016). Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 13–22).
33. Zhai, H., Lv, X., Hou, Z., Tong, X., & Bu, F. (2023). MLNet: A multi-level multimodal named entity recognition architecture. *Frontiers in Neurorobotics,17*, 1181143.
34. Zhang, H., Cai, Y., Ren, H., & Li, Q. (2023). Multimodal Topic Modeling by Exploring Characteristics of Short Text Social Media. *IEEE Transactions on Multimedia,25*, 2430–2445.
35. Zheng, L., Caiming, Z., & Caixian, C. (2018). MMDF-LDA: An improved multi-modal latent Dirichlet allocation model for social image annotation. *Expert Systems with Applications,104*, 168–184.