

Lexical Completion from Visual Evidence for Multimodal Topic Modeling

Dayu Guo^{1*} Zhiwen Luo^{1*} Nizar Bouguila² Wentao Fan^{1†}

¹Guangdong Provincial Key Laboratory IRADS and Department of
Computer Science, Beijing Normal-Hong Kong Baptist University

²Department of Cybersecurity and Intelligent Systems Engineering, Concordia University
lsdayuguo@bnbu.edu.cn, zhiwenluo@bnbu.edu.cn,
nizar.bouguila@concordia.ca, wentaofan@bnbu.edu.cn

Abstract

Multimodal topic modeling aims to learn interpretable topics from collections that combine text with other modalities. A persistent difficulty is lexical incompleteness: the text often records only part of the evidence needed to describe a sample, and concepts that are visible in images may have no direct path into the learned word distributions. In many multimodal topic models, visual features mainly help infer document-level topic mixtures, whereas the decoder is still trained on the incomplete observed text. We introduce TIGNTM, a graph completion framework that turns visual evidence into lexical supervision. TIGNTM builds a corpus graph over textual and visual representations, treats related samples as weak lexical witnesses, and produces sparse completion targets for words missing from the observed text. These targets supervise the same decoder that generates topic words, so visual cues can shape the word distributions that define each topic. To match the geometry of normalized multimodal embeddings, TIGNTM adopts hyperspherical neural topic learning. Experiments on multiple multimodal datasets show consistent improvements in coherence, diversity, and overall topic quality compared with strong neural and multimodal baselines.

1 Introduction

Topic models provide a compact and interpretable representation of large document collections by describing each document as a mixture of latent topics and each topic as a distribution over words (Blei et al., 2003). Neural topic models extend this paradigm with amortized inference and neural decoders, offering improved scalability and greater representational flexibility over classical probabilistic inference (Wu et al., 2024). Among them, Variational Autoencoder (VAE)-based neural topic models have become a dominant framework for topic

learning (Miao et al., 2017; Srivastava and Sutton, 2017; Luo et al., 2026b).

Recent advances have improved neural topic modeling from multiple perspectives. Contextualized topic models incorporate pretrained language representations to mitigate the sparsity and context-insensitivity of Bag-of-Words (BoW) inputs (Bianchi et al., 2021a,b). Graph-enhanced models exploit word- or document-level relations to improve semantic coherence (Shen et al., 2021; Zhu et al., 2024; Liu et al., 2025). In parallel, geometry-aware topic models have revisited the latent assumptions of neural topic models. Since many modern semantic embeddings are L_2 -normalized and compared mainly through angular similarity (Radford et al., 2021), hyperspherical latent spaces provide a more suitable alternative such as von Mises-Fisher (vMF) prior to Euclidean Gaussian assumptions (Xu et al., 2023; Guo et al., 2025; Adhya and Sanyal, 2025; Guo et al., 2026c). However, these developments have primarily focused on text-only corpora, leaving multimodal topic modeling less thoroughly explored.

Multimodal corpora, such as image captions, social media posts, and web documents, contain both textual and visual signals that often provide complementary semantic evidence (Zhang et al., 2022; Rashid et al., 2023). Existing multimodal neural topic models typically introduce visual or vision-language embeddings as auxiliary document-level features to improve topic inference (Gonzalez-Pizarro and Carenini, 2024; Abaskohi et al., 2025). While effective, this design still has two limitations. First, many methods retain Euclidean Gaussian latent assumptions, which are not well aligned with normalized multimodal embeddings whose semantics are largely directional (Li et al., 2025; Luo et al., 2025, 2026c,a). Second, visual information is usually used only to enhance the encoder, whereas the topic-word decoder is still trained mainly by reconstructing the observed text. Consequently, vi-

*Equal contribution.

†Corresponding author.



Figure 1: An example of lexical completion using lexical witnesses from visually similar samples.

visual concepts that are salient in an image but absent from its caption provide little direct supervision for learning topic-word distributions (Gonzalez-Pizarro and Carenini, 2024).

This exposes a central challenge in multimodal topic modeling: *lexical incompleteness*. An image may clearly contain certain objects or scenes, but surrounding texts often provide selective descriptions of such visual content and may even contain noise. When topic-word learning relies solely on observed words, these visually grounded but text-omitted concepts are difficult to recover, limiting both topic completeness and interpretability.

To address this challenge, we argue that a more effective multimodal topic model should not only encode visual features, but also transform visual evidence into lexical supervision for topic-word learning. As shown in Figure 1, in an image-text corpus, visually similar samples can provide weak lexical evidence for one another: words appearing in captions of visually related images may help infer concepts missing from a target caption. Graphs are well suited for propagating such cross-sample and cross-modal information (Gao et al., 2020; Li et al., 2023). However, existing graph-based neural topic models mainly use graphs to enhance representations, rather than to construct lexical completion targets for supervising topic-word distributions (Shen et al., 2021; Liu et al., 2025). Moreover, when graph edges are built from normalized text and image embeddings with cosine similarity, the propagated signals are guided by directional semantics (Zhu et al., 2024). This motivates a unified framework that combines graph-based lexical completion with hyperspherical topic inference.

Motivated by these observations, we propose the Text-Image Graph Neural Topic Model (TIGNTM), a graph-completed hyperspherical framework for multimodal topic modeling. Un-

like existing multimodal topic models that mainly use images as auxiliary encoder inputs, TIGNTM further converts visual and relational evidence into direct supervision for topic-word learning, leading to more complete and interpretable multimodal topics. Our contributions are summarized as follows:

- We identify *lexical incompleteness* as an important but underexplored limitation of multimodal neural topic models, and address it with a corpus-level text-image graph completion mechanism.
- We develop a graph-completed hyperspherical neural topic model that jointly supports directional multimodal topic inference and completion-aware topic-word learning.
- Experiments on multiple public multimodal datasets demonstrate that TIGNTM achieves better topic coherence, diversity, and overall topic quality than competitive neural, multimodal, and graph-enhanced baselines.

2 Related Work

Neural topic models. Classical topic models, represented by LDA, describe documents as mixtures of latent topics with Dirichlet priors (Blei et al., 2003). Neural topic models extend this framework using amortized inference and neural decoders, often with VAE-style objectives (Miao et al., 2017; Srivastava and Sutton, 2017). ETM improves topic-word estimation by embedding words and topics in a shared semantic space (Dieng et al., 2020), while contextualized topic models use pretrained language representations to alleviate the sparsity and context-insensitivity of BoW inputs (Bianchi et al., 2021a,b). Recent graph-enhanced methods exploit word- or document-level relations, such as dependency graphs or aligned contextual and graph-based topic signals, to improve topic coherence (Shen et al., 2021; Liu et al., 2025). In parallel, geometry-aware neural topic models replace Euclidean Gaussian priors with hyperspherical priors to better match normalized semantic embeddings and angular similarity structures (Xu et al., 2023; Adhya and Sanyal, 2025; Guo et al., 2026c). These studies improve topic representations from contextual, relational, or geometric perspectives, but most remain text-centered.

Multimodal topic modeling. Multimodal topic models aim to discover interpretable topics from

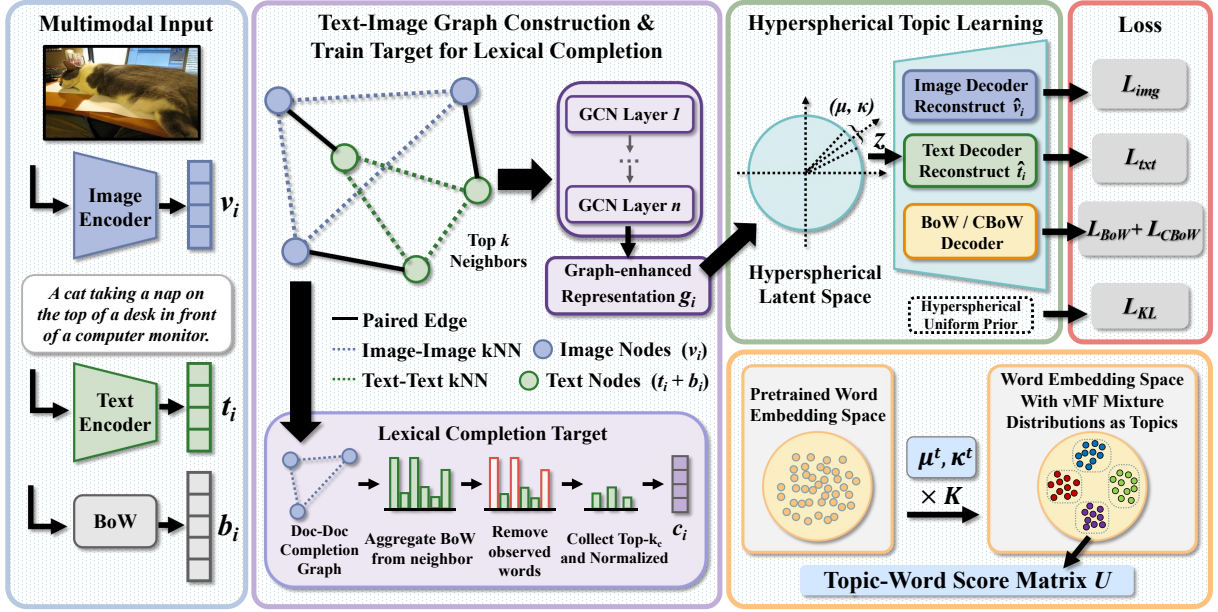


Figure 2: Overall architecture of TIGNTM. Paired images and texts are encoded as multimodal features. A corpus-level text-image graph is constructed to propagate neighbor information. The hyperspherical topic model then infers topics from graph-completed features and learns topic-word distributions through vMF mixture-based decoding.

paired or weakly aligned multimodal data. Early probabilistic models studied image annotation and image-text correspondence (Blei and Jordan, 2003), while later methods considered more realistic multimodal corpora such as social media posts and web documents, where textual and visual signals provide complementary semantic evidence (Zhang et al., 2022; Rashid et al., 2023). Recent neural multimodal topic models further incorporate visual features or pretrained vision-language embeddings as document-level inputs to improve topic inference (Zosa and Pivovarov, 2022; Gonzalez-Pizarro and Carenini, 2024; Abaskohi et al., 2025; Guo et al., 2026a). However, visual information is typically used mainly to enhance the encoder, whereas topic-word distributions are still learned from the observed text. Consequently, visually salient concepts that are absent from captions or posts provide limited direct supervision for topic-word learning. Unlike prior work, TIGNTM converts visual and relational evidence into lexical completion targets for the topic-word decoder, while using hyperspherical topic inference to model the directional geometry of normalized multimodal representations.

3 Methodology

Let $\mathcal{D} = \{(T_i, I_i)\}_{i=1}^N$ denote a corpus of N paired texts and images, and let V be the vocabulary size. Each document is represented by three comple-

mentary views: a sparse BoW vector $b_i \in \mathbb{R}^V$, a dense text embedding $t_i \in \mathbb{R}^{D_t}$, and a dense image embedding $v_i \in \mathbb{R}^{D_v}$. The key challenge is that b_i contains only words explicitly observed in T_i , whereas I_i may contain topically salient objects, scenes, or attributes that are absent from the text. Such lexical incompleteness limits the supervision available to topic-word learning.

To address this issue, we propose **TIGNTM**, a graph-completed hyperspherical neural topic model. Figure 2 shows the model overview. The model has two components. First, a text-image graph semantic completion module propagates cross-modal and cross-sample evidence to generate graph-completed document representations and visually grounded lexical completion targets. Second, a hyperspherical topic model infers document-topic distributions from BoW, text, image, and graph-completed features, using both observed and graph-completed words to supervise the same decoder.

3.1 Text-Image Graph Semantic Completion

The graph-completion module constructs a corpus-level text-image graph and derives two types of semantic completion signals: graph-completed document representations for topic inference and lexical completion targets for topic-word learning.

Graph construction. We construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}_{\mathcal{G}}, \mathcal{E}_{\mathcal{G}})$ with $2N$ nodes, $\mathcal{V}_{\mathcal{G}} =$

$\{u_i^t, u_i^v\}_{i=1}^N$, where u_i^t and u_i^v denote the text and image nodes of document i , respectively. Let $A \in \mathbb{R}^{2N \times 2N}$ denote the weighted adjacency matrix. Entries not associated with a retained edge are set to zero. Each paired text-image instance is first connected by bidirectional intra-pair edges:

$$A_{u_i^t, u_i^v} = A_{u_i^v, u_i^t} = \alpha_p, \quad (1)$$

where $\alpha_p \geq 0$ controls the contribution of paired cross-modal evidence.

We further define $\mathcal{N}_v(i)$ and $\mathcal{N}_t(i)$ as the sets of the k nearest image and text neighbors of document i , respectively. The neighbor sets are computed using cosine similarity between the corresponding text or image embeddings. The retained image-image and text-text edges are assigned weights

$$A_{u_i^v, u_j^v} = A_{u_j^v, u_i^v} = \alpha_v \cos(\mathbf{v}_i, \mathbf{v}_j), \quad j \in \mathcal{N}_v(i), \quad (2)$$

$$A_{u_i^t, u_j^t} = A_{u_j^t, u_i^t} = \alpha_t \cos(\mathbf{t}_i, \mathbf{t}_j), \quad j \in \mathcal{N}_t(i), \quad (3)$$

where $\alpha_v \geq 0$ and $\alpha_t \geq 0$ scale the contributions of visual and textual neighbor edges. Edges not exceeding the similarity threshold are removed, and each retained edge is symmetrized by adding its reverse with the same weight.

Let I_{2N} be the $2N$ -dimensional identity matrix and let $D \in \mathbb{R}^{2N \times 2N}$ be the diagonal degree matrix of $A + I_{2N}$, with $D_{rr} = \sum_s (A + I_{2N})_{rs}$, where $r, s \in \{1, \dots, 2N\}$ index the graph nodes. After adding self-loops, the adjacency matrix is symmetrically normalized as

$$\tilde{A} = D^{-1/2}(A + I_{2N})D^{-1/2}. \quad (4)$$

Thus, the resulting normalized adjacency matrix jointly encodes paired cross-modal relations, within-modality semantic neighbors, and self-information.

Graph encoder. Let D_g denote the graph hidden dimension. The initial text-node representation combines lexical statistics with contextual text semantics:

$$\mathbf{h}_i^{t,0} = \text{LN}(W_b \mathbf{b}_i + W_t \mathbf{t}_i), \quad (5)$$

where $W_b \in \mathbb{R}^{D_g \times V}$ and $W_t \in \mathbb{R}^{D_g \times D_t}$ are learnable projection matrices for the BoW and text representations, respectively, and $\text{LN}(\cdot)$ denotes layer normalization. The corresponding image node is initialized as

$$\mathbf{h}_i^{v,0} = \text{LN}(W_v \mathbf{v}_i), \quad (6)$$

where $W_v \in \mathbb{R}^{D_g \times D_v}$ is the learnable image projection matrix.

Let $H^\ell \in \mathbb{R}^{2N \times D_g}$ denote the matrix whose rows contain all text- and image-node representations at graph layer ℓ . Using the initialized text and image node representations, we apply L residual graph-convolution layers (Kipf and Welling, 2017):

$$H^{\ell+1} = H^\ell + \text{Dropout}\left(\sigma\left(\tilde{A}H^\ell W^\ell\right)\right), \quad (7)$$

where $W^\ell \in \mathbb{R}^{D_g \times D_g}$ is the learnable graph-convolution weight at layer ℓ , and $\sigma(\cdot)$ denotes the nonlinear activation function. The residual connection preserves the original node information while the multiplication by $\tilde{A}$ aggregates evidence from paired and neighboring nodes.

After graph propagation, the final text and image states of document i are fused as

$$\mathbf{g}_i = \text{norm}\left(f_g\left([h_i^{t,L}; h_i^{v,L}]\right)\right), \quad (8)$$

where $f_g : \mathbb{R}^{2D_g} \rightarrow \mathbb{R}^{D_g}$ is a learnable fusion network and $\text{norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. Unlike a direct concatenation of raw multimodal features, $\mathbf{g}_i$ incorporates information from the paired modality and semantically related corpus samples.

Lexical completion targets. To convert visual evidence into lexical supervision, we derive a document-level completion graph S from the image-image subgraph of A . Let $S \in \mathbb{R}^{N \times N}$ denote its row-normalized weighted adjacency matrix, where $S_{ij} > 0$ only if documents i and j are retained as visual neighbors, and $\sum_j S_{ij} = 1$ for every node with at least one neighbor. Before row normalization, the edge weights are given by the cosine similarities between image embeddings.

Let $B \in \mathbb{R}^{N \times V}$ denote the observed BoW matrix whose i -th row is $\mathbf{b}_i^\top$. The lexical evidence collected from visually related documents is

$$\mathbf{m}_i^\top = S_i B = \sum_{j=1}^N S_{ij} \mathbf{b}_j^\top, \quad (9)$$

where S_i is the i -th row of S . A word receives a large value in $\mathbf{m}_i$ when it frequently occurs in captions associated with visually similar samples.

Because the completion target should contain only missing evidence, words already observed in the target document are removed:

$$\bar{m}_{iv} = m_{iv} \cdot \mathbb{I}[b_{iv} = 0], \quad (10)$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Let Ω_i be the indices of the top K_c positive entries in $\bar{\mathbf{m}}_i$. The sparse lexical completion vector is defined as

$$c_{iv} = \begin{cases} \bar{m}_{iv} / \max_{u \in \Omega_i} \bar{m}_{iu}, & v \in \Omega_i, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Thus, $\mathbf{c}_i \in [0, 1]^V$ represents lexical evidence that is absent from the paired text but supported by visually similar samples. It provides weak yet direct supervision for visually salient concepts that would otherwise be invisible to the topic-word decoder.

Graph-completion learning. To make the graph encoder robust to missing or incomplete modality information, we randomly mask text and image embeddings of mini-batch documents during graph-completion learning. The graph is still propagated over the corpus, encouraging each node to recover missing evidence from its paired modality and neighboring samples. The graph-completion objective is

$$\mathcal{L}_{gc} = \mathcal{L}_{\text{pair}} + \lambda_r (\mathcal{L}_{\text{image}} + \mathcal{L}_{\text{text}}) + \lambda_{\text{comp}} \mathcal{L}_{\text{comp}}, \quad (12)$$

where $\mathcal{L}_{\text{pair}}$ is a symmetric InfoNCE loss between paired propagated text and image states (Oord et al., 2018), $\mathcal{L}_{\text{image}}$ and $\mathcal{L}_{\text{text}}$ reconstruct the original image and text embeddings from $\mathbf{g}_i$, and $\mathcal{L}_{\text{comp}}$ is the binary cross-entropy loss, encouraging $\mathbf{g}_i$ to encode lexical completion evidence. The coefficient $\lambda_r \geq 0$ controls the joint contribution of the two modality-reconstruction losses, and $\lambda_{\text{comp}} \geq 0$ controls the contribution of lexical completion prediction.

After graph-completion training, the representations and targets $\{(\mathbf{g}_i, \mathbf{c}_i)\}_{i=1}^N$ are supplied to the downstream topic model. The former improves document-level topic inference, whereas the latter provides direct lexical supervision to the topic-word decoder.

3.2 Graph-Completed Hyperspherical Topic Model

Given the graph-completed representation $\mathbf{g}_i$ and lexical completion target $\mathbf{c}_i$, TIGNTM infers document-topic distributions in a hyperspherical latent space. The encoder input integrates observed lexical evidence, raw multimodal features, and graph-completed representations:

$$\mathbf{x}_i = [\mathbf{b}_i; P_g \mathbf{g}_i; P_t \mathbf{t}_i; P_v \mathbf{v}_i], \quad (13)$$

where P_g , P_t , and P_v are learnable projection layers. The inference network maps $\mathbf{x}_i$ to a vMF posterior $q(\mathbf{z}_i | \mathbf{x}_i) = \text{vMF}(\boldsymbol{\mu}_i, \kappa_i)$ on the unit hypersphere, where $\boldsymbol{\mu}_i$ is the mean direction and κ_i controls posterior concentration. We use the uniform distribution on the hypersphere as the prior $p(\mathbf{z})$. A latent code is sampled via a differentiable vMF reparameterization (Davidson et al., 2018):

$$\mathbf{z}_i \sim q(\mathbf{z}_i | \mathbf{x}_i), \quad \boldsymbol{\theta}_i = \text{softmax}(\tau \cdot \mathbf{z}_i), \quad (14)$$

where $\boldsymbol{\theta}_i$ denotes the document-topic distribution. The vMF reparameterization and KL computation are provided in Appendix A and Appendix B, respectively. This hyperspherical formulation is consistent with the directional geometry of normalized text, image, and graph-completed representations.

vMF Mixture topic-word decoder. To obtain interpretable topics, each topic is represented as a vMF distribution in the word embedding space. Let $\mathbf{w}_v \in \mathbb{R}^{D_e}$ denote the embedding of word v . For topic k , we learn a topic direction $\boldsymbol{\mu}_k^t$ and a concentration parameter κ_k^t . The topic-word compatibility is computed by the vMF density f_{vMF} :

$$\begin{aligned} u_{kv} &= \log f_{\text{vMF}}(\boldsymbol{\mu}_k^t, \kappa_k^t, \mathbf{w}_v) \\ &= \log C_{D_e}(\kappa_k^t) + \kappa_k^t (\boldsymbol{\mu}_k^t)^\top \mathbf{w}_v, \end{aligned} \quad (15)$$

where $C_{D_e}(\cdot)$ is the normalization constant.

This score corresponds to the vMF log-density up to a topic-specific constant, which cancels out when applying softmax over the vocabulary. The topic-word distribution is therefore

$$\boldsymbol{\beta}_k = \text{softmax}([u_{k1}, \dots, u_{kV}]^\top). \quad (16)$$

For document reconstruction, the topic mixture is mapped to a word distribution $\hat{\mathbf{b}}_i = \text{softmax}(\boldsymbol{\theta}_i U)$, where $U = [u_{kv}] \in \mathbb{R}^{K \times V}$.

To preserve multimodal semantics in the latent space, the same hyperspherical code is also used to reconstruct dense text and image embeddings:

$$\hat{\mathbf{t}}_i = \mathbf{z}_i B_t, \quad \hat{\mathbf{v}}_i = \mathbf{z}_i B_v. \quad (17)$$

These auxiliary reconstruction branches encourage the latent topics to retain multimodal information, while the topic-word decoder remains responsible for producing interpretable lexical topics.

3.3 Completion-Aware Topic Objective

The observed BoW vector and the graph-derived completion vector supervise the same decoder distribution $\hat{\mathbf{b}}_i$. The standard BoW reconstruction loss

and the completion-aware lexical supervision are

$$\mathcal{L}_{\text{bow}}(i) = -\mathbf{b}_i^\top \log \hat{\mathbf{b}}_i, \quad \mathcal{L}_{\text{cbow}}(i) = -\mathbf{c}_i^\top \log \hat{\mathbf{b}}_i. \quad (18)$$

The full topic learning objective is

$$\mathcal{L}_{\text{topic}} = \frac{1}{N} \sum_{i=1}^N \left(\mathcal{L}_{\text{bow}}(i) + \lambda_{\text{cbow}} \mathcal{L}_{\text{cbow}}(i) + \mathcal{L}_{\text{img}}(i) + \mathcal{L}_{\text{txt}}(i) + \mathcal{L}_{\text{KL}}(i) \right), \quad (19)$$

where $\mathcal{L}_{\text{txt}}$ and $\mathcal{L}_{\text{img}}$ are cosine reconstruction losses for text and image embeddings, and $\mathcal{L}_{\text{KL}}$ is the KL divergence between the vMF posterior $q(\mathbf{z}_i | \mathbf{x}_i)$ and the hyperspherical uniform prior $p(\mathbf{z})$. We set λ_{cbow} smaller than the observed BoW weight so that the original text remains the primary lexical evidence, while graph-derived completion words serve as weak supervision.

This objective closes the semantic-completion loop: the text-image graph identifies visually supported missing lexical evidence, and the topic model directly absorbs this evidence through completion-aware BoW supervision rather than using graph information only as an auxiliary encoder feature.

4 Experiments

4.1 Datasets and Preprocessing

We evaluate TIGNTM on three public image-text datasets that cover different multimodal topic modeling scenarios. Flickr30k (Young et al., 2014) contains descriptive image captions, COCO2014 (Lin et al., 2014) covers diverse everyday scenes with human-written captions, and Twitter100k (Hu et al., 2018) consists of short and noisy social media posts. These datasets allow us to evaluate topic modeling under both well-described caption settings and sparse, informal text settings. Dataset statistics are summarized in Table 1.

Table 1: Dataset statistics.

Dataset	# Documents	Avg. length	Vocab size
Flickr30k	31,014	18.0	2,174
Twitter100k	100,000	7.4	8,092
COCO2014	82,783	10.5	2,065

We use the same preprocessing pipeline for all datasets. Texts are tokenized and lemmatized, stopwords are removed, and the vocabulary is filtered by document frequency: words appearing in fewer than 10 documents or in more than 5% of documents are discarded. Text and image embeddings

are extracted using *CLIP-ViT-B/32* to ensure all multimodal methods are evaluated under the same feature extraction setting whenever applicable.

4.2 Baselines and Implementation

We compare TIGNTM with representative baselines from five categories: classical/neural topic models, including LDA (Blei et al., 2003), NeuralLDA (Srivastava and Sutton, 2017), and ETM (Dieng et al., 2020); contextual topic models, including ZeroShotTM (Bianchi et al., 2021b), CombinedTM (Bianchi et al., 2021a), and BERTopic (Grootendorst, 2022); geometry-aware topic models, including vONT (Xu et al., 2023), NVGMTM (Tang et al., 2023), S2WTM (Adhya and Sanyal, 2025), and vMN-VTM (Guo et al., 2026c); multimodal topic models, including Multimodal-ZeroShotTM (Gonzalez-Pizarro and Carenini, 2024), Multimodal-Contrast (Gonzalez-Pizarro and Carenini, 2024), and CEMTM (Abaskohi et al., 2025); and graph-enhanced topic models, including CGTM (Liu et al., 2025). More implementation details about our model and baseline shown in the Appendix C.

We evaluate topic quality from the complementary perspectives of coherence and diversity. NPMI and C_v measure the semantic coherence of top topic words (Röder et al., 2015). Topic Diversity (TD) and Inverted Rank-Biased Overlap (I-RBO) measure topic distinctiveness and redundancy across learned topics (Webber et al., 2010; Terragni et al., 2021). Following prior work (Dieng et al., 2020; Guo et al., 2026b,d), we also report an overall Quality score, computed as $C_v \times \text{TD}$, summarizing the trade-off between coherent and non-redundant topics. Higher values indicate better performance for all five metrics.

4.3 Main Results

Table 2 reports the quantitative results across three datasets and different topic numbers. TIGNTM consistently achieves the best NPMI, C_v , and overall Quality scores in most settings, demonstrating its advantage in learning coherent and diverse multimodal topics. While some baselines obtain competitive TD or I-RBO scores, their coherence scores are often lower, indicating that topic distinctiveness alone does not necessarily yield semantically meaningful topics.

The improvements of TIGNTM directly support our main motivation. Existing multimodal topic models such as Multimodal-ZeroShotTM typ-

Table 2: Topic quality on COCO2014, Twitter100k, and Flickr30k. All metrics are computed using the top-10 words of each topic. Within each cell, the three values correspond to $K = 10, 20$, and 50 topics, respectively. Red and blue indicate the best and second-best results for each metric and topic number within each dataset.

Model	NPMI ($K = 10/20/50$)	C_v ($K = 10/20/50$)	TD ($K = 10/20/50$)	I-RBO ($K = 10/20/50$)	Quality ($K = 10/20/50$)
COCO2014					
LDA (2003)	0.042 / 0.044 / -0.021	0.430 / 0.425 / 0.385	0.890 / 0.842 / 0.763	0.957 / 0.977 / 0.979	0.383 / 0.358 / 0.294
NeuralLDA (2017)	-0.124 / -0.081 / -0.081	0.334 / 0.347 / 0.364	1.000 / 0.968 / 0.826	0.999 / 1.000 / 0.995	0.334 / 0.336 / 0.301
ETM (2020)	-0.084 / -0.064 / -0.079	0.240 / 0.234 / 0.227	0.120 / 0.085 / 0.063	0.181 / 0.302 / 0.472	0.029 / 0.020 / 0.014
ZeroShotTM (2021)	-0.075 / -0.015 / -0.082	0.373 / 0.442 / 0.366	0.910 / 0.720 / 0.323	0.990 / 0.976 / 0.937	0.339 / 0.318 / 0.118
CombinedTM (2021)	-0.064 / -0.059 / -0.003	0.379 / 0.394 / 0.401	0.980 / 0.790 / 0.390	0.996 / 0.985 / 0.955	0.371 / 0.311 / 0.156
BERTopic (2022)	0.064 / 0.110 / 0.129	0.441 / 0.471 / 0.506	0.860 / 0.778 / 0.686	0.967 / 0.967 / 0.985	0.379 / 0.366 / 0.347
vONT (2023)	-0.099 / -0.059 / -0.119	0.361 / 0.353 / 0.354	0.998 / 0.970 / 0.989	1.000 / 0.990 / 0.999	0.360 / 0.342 / 0.350
NVGMTM (2023)	-0.113 / -0.084 / -0.054	0.247 / 0.278 / 0.295	0.730 / 0.551 / 0.347	0.951 / 0.948 / 0.940	0.180 / 0.153 / 0.102
Multimodal-ZeroShotTM (2024)	-0.076 / -0.028 / -0.035	0.318 / 0.420 / 0.381	0.970 / 0.780 / 0.364	0.995 / 0.981 / 0.957	0.308 / 0.328 / 0.139
S2WTM (2025)	-0.328 / -0.362 / -0.244	0.432 / 0.414 / 0.420	0.790 / 0.891 / 0.860	0.946 / 0.992 / 0.994	0.341 / 0.369 / 0.361
CEMTM (2025)	0.072 / 0.082 / 0.063	0.446 / 0.466 / 0.445	0.981 / 0.944 / 0.777	0.989 / 0.984 / 0.981	0.438 / 0.440 / 0.346
CGTM (2025)	0.081 / 0.059 / 0.074	0.499 / 0.411 / 0.441	0.887 / 0.917 / 0.874	0.993 / 0.992 / 0.970	0.443 / 0.377 / 0.385
vMNVMTM (2026)	0.055 / 0.119 / 0.163	0.438 / 0.531 / 0.606	0.960 / 0.840 / 0.686	0.997 / 0.986 / 0.985	0.420 / 0.446 / 0.416
Ours	0.084 / 0.089 / 0.071	0.471 / 0.503 / 0.470	0.960 / 0.930 / 0.665	0.996 / 0.996 / 0.957	0.452 / 0.468 / 0.313
	0.085 / 0.166 / 0.183	0.535 / 0.616 / 0.622	1.000 / 0.970 / 0.884	1.000 / 0.998 / 0.996	0.535 / 0.598 / 0.550
Twitter100k					
LDA (2003)	0.007 / -0.042 / -0.142	0.370 / 0.305 / 0.269	0.850 / 0.967 / 0.961	0.965 / 0.998 / 0.998	0.315 / 0.295 / 0.259
NeuralLDA (2017)	-0.255 / -0.197 / -0.221	0.400 / 0.332 / 0.381	1.000 / 0.994 / 0.950	1.000 / 0.997 / 0.998	0.400 / 0.330 / 0.362
ETM (2020)	0.004 / 0.005 / -0.028	0.360 / 0.359 / 0.378	0.210 / 0.185 / 0.172	0.299 / 0.426 / 0.497	0.076 / 0.066 / 0.065
ZeroShotTM (2021)	-0.175 / -0.264 / -0.364	0.409 / 0.407 / 0.450	0.960 / 0.678 / 0.470	0.992 / 0.964 / 0.958	0.393 / 0.276 / 0.211
CombinedTM (2021)	-0.159 / -0.252 / -0.383	0.401 / 0.435 / 0.465	0.950 / 0.668 / 0.512	0.986 / 0.958 / 0.963	0.381 / 0.291 / 0.238
BERTopic (2022)	0.027 / 0.061 / -0.033	0.411 / 0.486 / 0.382	0.920 / 0.920 / 0.960	0.980 / 0.994 / 0.998	0.378 / 0.447 / 0.367
vONT (2023)	-0.011 / -0.060 / -0.148	0.411 / 0.331 / 0.333	0.980 / 1.000 / 1.000	0.995 / 1.000 / 1.000	0.403 / 0.331 / 0.333
NVGMTM (2023)	-0.028 / -0.030 / -0.003	0.384 / 0.330 / 0.346	0.690 / 0.510 / 0.380	0.913 / 0.875 / 0.876	0.265 / 0.168 / 0.131
Multimodal-ZeroShotTM (2024)	-0.286 / -0.239 / -0.361	0.381 / 0.378 / 0.431	0.950 / 0.780 / 0.500	0.992 / 0.977 / 0.963	0.362 / 0.295 / 0.215
S2WTM (2025)	-0.436 / -0.423 / -0.418	0.501 / 0.507 / 0.505	0.850 / 0.900 / 0.828	0.978 / 0.991 / 0.993	0.426 / 0.456 / 0.418
CEMTM (2025)	-0.045 / -0.053 / -0.041	0.313 / 0.356 / 0.368	0.955 / 0.925 / 0.929	0.916 / 0.916 / 0.951	0.299 / 0.329 / 0.342
CGTM (2025)	0.017 / 0.023 / 0.004	0.442 / 0.416 / 0.390	0.972 / 0.990 / 0.919	0.903 / 0.879 / 0.900	0.430 / 0.412 / 0.358
vMNVMTM (2026)	0.038 / 0.056 / 0.050	0.437 / 0.454 / 0.441	0.740 / 0.668 / 0.519	0.926 / 0.954 / 0.960	0.323 / 0.303 / 0.229
Ours	0.030 / 0.016 / -0.089	0.426 / 0.431 / 0.346	0.985 / 0.960 / 0.930	0.996 / 0.995 / 0.996	0.420 / 0.414 / 0.322
	0.101 / 0.111 / 0.112	0.523 / 0.567 / 0.551	1.000 / 0.995 / 0.916	1.000 / 0.999 / 0.993	0.523 / 0.564 / 0.505
Flickr30k					
LDA (2003)	0.040 / 0.021 / -0.027	0.459 / 0.428 / 0.383	0.570 / 0.608 / 0.680	0.850 / 0.912 / 0.949	0.262 / 0.260 / 0.260
NeuralLDA (2017)	-0.148 / -0.142 / -0.155	0.324 / 0.289 / 0.309	0.980 / 0.930 / 0.690	0.977 / 0.997 / 0.989	0.318 / 0.269 / 0.213
ETM (2020)	-0.001 / -0.009 / -0.033	0.372 / 0.361 / 0.385	0.190 / 0.134 / 0.112	0.232 / 0.313 / 0.499	0.071 / 0.048 / 0.043
ZeroShotTM (2021)	-0.073 / -0.048 / -0.048	0.446 / 0.487 / 0.491	0.860 / 0.609 / 0.335	0.957 / 0.945 / 0.915	0.384 / 0.297 / 0.164
CombinedTM (2021)	-0.031 / -0.003 / -0.055	0.476 / 0.481 / 0.495	0.770 / 0.620 / 0.326	0.950 / 0.949 / 0.917	0.367 / 0.298 / 0.161
BERTopic (2022)	0.033 / 0.080 / 0.075	0.445 / 0.500 / 0.515	0.670 / 0.720 / 0.725	0.906 / 0.956 / 0.987	0.298 / 0.360 / 0.373
vONT (2023)	-0.026 / -0.055 / -0.189	0.370 / 0.325 / 0.311	0.940 / 1.000 / 0.984	0.995 / 1.000 / 0.999	0.348 / 0.325 / 0.307
NVGMTM (2023)	-0.028 / -0.006 / -0.014	0.282 / 0.322 / 0.299	0.680 / 0.488 / 0.264	0.941 / 0.926 / 0.915	0.192 / 0.157 / 0.079
Multimodal-ZeroShotTM (2024)	-0.004 / -0.006 / 0.001	0.462 / 0.455 / 0.446	0.870 / 0.700 / 0.337	0.980 / 0.970 / 0.936	0.402 / 0.319 / 0.150
S2WTM (2025)	-0.253 / -0.228 / -0.219	0.343 / 0.445 / 0.429	0.720 / 0.793 / 0.673	0.943 / 0.982 / 0.981	0.247 / 0.353 / 0.289
CEMTM (2025)	0.058 / 0.071 / 0.056	0.489 / 0.513 / 0.515	0.811 / 0.757 / 0.629	0.882 / 0.897 / 0.924	0.397 / 0.388 / 0.324
CGTM (2025)	-0.008 / -0.015 / 0.006	0.457 / 0.405 / 0.458	0.955 / 0.917 / 0.827	0.931 / 0.927 / 0.860	0.436 / 0.371 / 0.379
vMNVMTM (2026)	0.064 / 0.079 / 0.080	0.473 / 0.485 / 0.499	0.740 / 0.690 / 0.523	0.955 / 0.970 / 0.971	0.350 / 0.335 / 0.261
Ours	0.036 / 0.068 / -0.016	0.442 / 0.492 / 0.386	0.980 / 0.980 / 0.800	0.991 / 0.991 / 0.987	0.433 / 0.482 / 0.309
	0.117 / 0.153 / 0.172	0.681 / 0.702 / 0.712	0.985 / 0.910 / 0.784	0.998 / 0.998 / 0.991	0.671 / 0.639 / 0.558

ically use images as auxiliary document-level features, leaving topic-word learning dominated by observed text. In contrast, TIGNTM converts visual and relational evidence into lexical completion targets, which provide additional weak supervision to the topic-word decoder. Combined with hyper-spherical topic inference, this design improves the coherence-diversity trade-off and leads to more complete topic-word distributions.

4.4 Qualitative Analysis

To examine whether TIGNTM recovers visually grounded evidence absent from paired text, Figure 3 presents representative completion examples. These examples show that visually related samples provide lexical cues for concepts absent from the target text. These graph-derived completion words enrich topic-word decoder supervision and capture visual semantics beyond the observed caption.

We also provide other qualitative examples of the multimodal topics learned by TIGNTM on differ-

Case: Skateboard Trick (Flickr30k)



Original caption: "A boy is riding a skateboard down a slope that leads from a bridge to the street below."

Visually completed words:
ramp, trick, midair

Final topic words:
ramp, trick, tricks, skateboarder, stunt, midair, skate, skateboard, skateboarding, rail

Case: Political Speech (Twitter100k)



Original caption: "Donald Trump endorses House Speaker Paul Ryan."

Visually completed words:
speech, president

Final topic words:
trump, hillary, donald, clinton, hillaryclinton, realdonaldtrump, obama, **president**, demsinphilly, **speech**

Figure 3: The examples introduce topic words such as *speech/president* and *ramp/trick/midair*, which are suggested by visually similar samples and not explicitly mentioned in the paired texts.

ent datasets in Figures 4–6. For each selected topic, we report its top keywords and the five image-text

pairs with the highest inferred topic probabilities. These examples complement the quantitative evaluation by showing whether the learned topic words are aligned with visually and textually coherent document groups across caption-style and social-media-style corpora.



Figure 4: Visualization examples of multimodal topic samples in COCO2014.

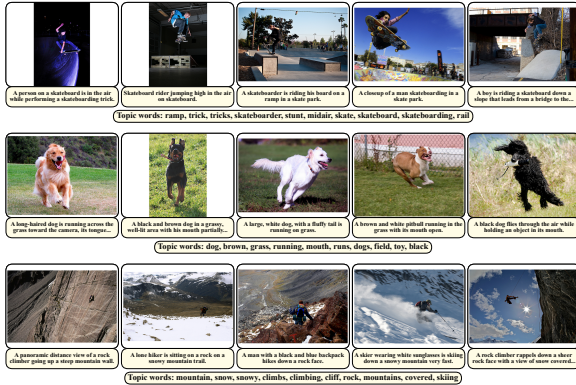


Figure 5: Visualization examples of multimodal topic samples in Flickr30k.

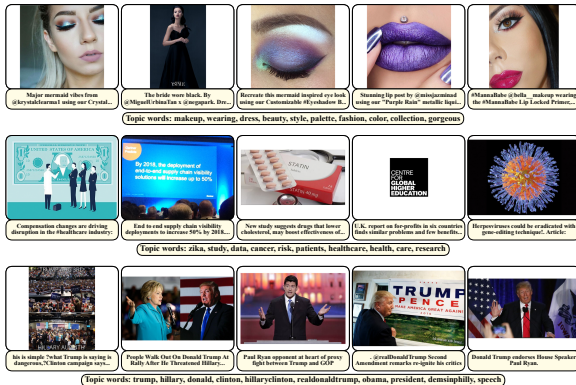


Figure 6: Visualization examples of multimodal topic samples in Twitter100k.

We further visualize the learned document repre-

sentations on Twitter100k using t-SNE. Each document is labeled by its most probable topic. As shown in Figure 7, documents assigned to the same dominant topic form compact regions, while different topics are relatively well separated. This suggests that graph-based lexical completion and hyperspherical topic inference help organize noisy multimodal documents into more coherent topic-aware representations.

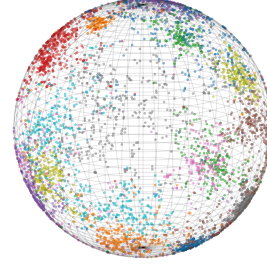


Figure 7: t-SNE visualization of document representations learned by TIGNTM on Twitter100k (20 topics). Each point denotes a document, and colors indicate the dominant inferred topics.

4.5 Ablation Study

We conduct ablation studies to assess the contribution of each major component. The variants remove: 1) graph-completed document representations, 2) the completion-aware BoW objective, 3) raw text and image encoder inputs, and 4) the hyperspherical prior and vMF topic-word decoder.

Table 3: Ablation results on representative topic settings. All metrics are computed from the top-10 topic words.

Variant	NPMI	C_v	TD	I-RBO	Quality
<i>COCO2014</i> , $K = 10$					
Full model	0.085	0.535	1.000	1.000	0.535
w/o raw text/image inputs	0.062	0.528	0.990	0.999	0.523
w/o completion BoW	0.057	0.517	0.995	0.998	0.514
w/o graph input	0.059	0.490	0.985	0.990	0.482
w/o hyperspherical topic learning	0.029	0.433	0.970	0.997	0.420
<i>Flickr30k</i> , $K = 20$					
Full model	0.153	0.702	0.910	0.998	0.639
w/o raw text/image inputs	0.134	0.667	0.875	0.992	0.584
w/o completion BoW	0.132	0.672	0.855	0.990	0.575
w/o graph input	0.143	0.681	0.870	0.991	0.592
w/o hyperspherical topic learning	0.067	0.566	0.505	0.744	0.286
<i>Twitter100k</i> , $K = 20$					
Full model	0.111	0.567	0.995	0.999	0.564
w/o raw text/image inputs	0.061	0.501	0.969	0.978	0.486
w/o completion BoW	0.062	0.499	0.970	0.998	0.484
w/o graph input	0.051	0.484	0.985	0.999	0.477
w/o hyperspherical topic learning	-0.009	0.324	0.380	0.829	0.123

As shown in Table 3, the full model achieves the best overall performance across all datasets. Removing graph-completed representations weakens document-level topic inference, while removing the completion-aware BoW objective reduces the

model’s ability to inject visually grounded lexical evidence into topic-word learning. These results indicate that the two graph-derived signals are complementary: graph-completed representations improve the encoder, whereas lexical completion targets directly supervise the decoder. Removing raw text and image inputs also degrades performance, showing that pretrained multimodal embeddings provide useful evidence beyond graph propagation. The largest drops are observed when replacing hyperspherical topic learning, especially on Flickr30k and Twitter100k, supporting the importance of modeling normalized multimodal representations with directional geometry.

4.6 Hyperparameter Analysis

We analyze the graph-completion module’s sensitivity to completion size K_c and completion-aware BoW weight λ_{cbow} . Results appear in Figure 8. The model is relatively stable across K_c values. Increasing K_c from 16 to 128 causes only minor metric fluctuations, suggesting TIGNTM does not rely on a precisely tuned number of completion words. In contrast, λ_{cbow} has a clearer effect on performance. A moderate completion weight best balances observed text supervision and graph-derived lexical completion. At $\lambda_{cbow} = 0.5$, the model achieves the best C_v , Diversity, and Quality scores, demonstrating the effectiveness of the completion BoW objective. Overall, the hyperparameter analysis shows that graph-based lexical completion provides robust additional supervision and consistently improves topic quality.

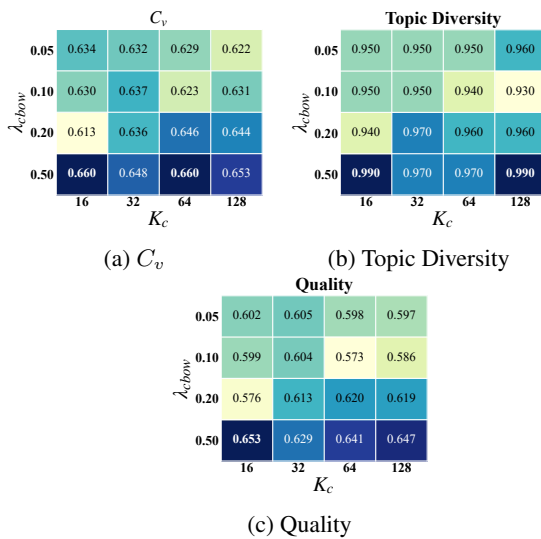


Figure 8: Hyperparameter sensitivity under different completion settings on Flickr30k with 10 topics.

4.7 Runtime Report

We report the training time across multimodal neural topic models in Figure 9 to assess computational cost. For each method, the total time is computed from its full training epochs. For TIGNTM, we additionally include the graph-completion time before topic learning. Although this stage introduces extra preprocessing overhead, TIGNTM remains competitive across multimodal baselines. Notably, CEMTM, a recent state-of-the-art multimodal topic model, incurs substantially higher computational cost, indicating limited training efficiency despite its strong modeling capacity.

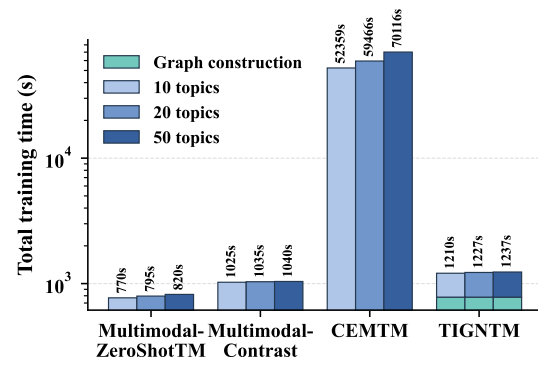


Figure 9: Training time on COCO2014 across topic numbers. Each bar reports total runtime in seconds. For TIGNTM, the green segment denotes graph-completion cost and the upper segment topic-learning time.

5 Conclusion

In this paper, we proposed TIGNTM, a graph-completed hyperspherical neural topic model for multimodal topic modeling. Motivated by the lexical incompleteness of paired text-image corpora, TIGNTM goes beyond using images as auxiliary encoder inputs and converts visual and relational evidence into lexical completion targets for topic-word learning. Specifically, it constructs a corpus-level text-image graph to propagate cross-modal and cross-sample semantic evidence, producing graph-completed document representations and visually grounded completion signals. These signals are integrated with a hyperspherical topic model that better matches the directional geometry of normalized multimodal representations. Experiments on multiple public text-image datasets demonstrate that TIGNTM improves topic coherence, diversity, and overall topic quality over competitive neural, multimodal, and graph-enhanced baselines.

Limitations

Although TIGNTM improves multimodal topic quality, it has several limitations. First, the construction of the corpus-level text-image graph depends on the quality of pretrained text and image embeddings. If these embeddings fail to capture reliable semantic similarity, the generated lexical completion targets may introduce noisy supervision. Second, the graph construction and propagation introduce additional computational cost compared with standard neural topic models, especially for large-scale corpora. Future work may explore more efficient graph sparsification or retrieval-based completion strategies to improve scalability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (W2632076, 62276106), Guangdong Basic and Applied Basic Research Foundation (2024A1515011767), and the Guangdong Provincial Key Laboratory IRADS (2022B1212010006).

References

- Amirhossein Abaskohi, Raymond Li, Chuyuan Li, Shafiq Joty, and Giuseppe Carenini. 2025. CEMTM: Contextual embedding-based multimodal topic modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11675–11692.
- Suman Adhya and Debarshi Kumar Sanyal. 2025. S2WTM: Spherical sliced-wasserstein autoencoder for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23211–23225.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683.
- David M Blei and Michael I Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. 2018. Hyperspherical variational auto-encoders. In *UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI).
- Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.
- Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12743–12753. IEEE.
- Felipe Gonzalez-Pizarro and Giuseppe Carenini. 2024. Neural multimodal topic modeling: A comprehensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12159–12172.
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. 2025. Biovmnvtm: A geometry-aware neural topic model for biomedical text analysis via von mises–fisher mixtures. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3648–3653. IEEE.
- Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. 2026a. MSW-NTM: A spherical wasserstein autoencoder for multimodal neural topic modeling with llm-guided topic refinement. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3759–3763.
- Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. 2026b. Multimodal topic discovery in web media via von mises-fisher mixture neural topic models. In *Proceedings of the ACM Web Conference 2026*, pages 8749–8752.
- Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. 2026c. Neural topic modeling on hyperspheres: Spherical representation learning with von mises–fisher mixtures. *Neural Networks*, 200:108792.
- Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. 2026d. ST-HNTM: Joint speech-text neural topic modeling on the hypersphere. In *ICASSP*

- 2026-2026 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12852–12856. IEEE.
- Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2018. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938.
- Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.
- Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427.
- Zhixiang Li, Zhiwen Luo, Nizar Bouguila, Weifeng Su, and Wentao Fan. 2025. Disentangled representation learning for multi-view clustering via von mises–fisher hyperspherical embedding. *Neural Networks*, 191:107802.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.
- Jiyuan Liu, Jiaxing Yan, Chunjiang Zhu, Xingyu Liu, Qing Li, and Yanghui Rao. 2025. Neural topic modeling via contextual and graph information fusion. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13247–13263.
- Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. 2025. Dynamic deep clustering of high-dimensional directional data via hyperspherical embeddings with bayesian nonparametric mixtures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 938–949.
- Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. 2026a. Adaptive deep clustering via disentangled hyperspherical vaes with contrastive geometry optimization. *Neurocomputing*, page 132625.
- Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. 2026b. Flexible priors for cross-collection topic modeling: an approach with generalized dirichlet and beta-liouville distributions. *International Journal of Data Science and Analytics*, 22(1):83.
- Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. 2026c. Hyperspherical representation learning of axial data via axial vaes with watson distribution. *ACM Transactions on Knowledge Discovery from Data*, 20(4):1–21.
- Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2410–2419.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.
- Junaid Rashid, Jungeun Kim, and Usman Naseem. 2023. Coherent topic modeling for creative multimodal data on social media. In *Proceedings of the ACM Web Conference 2023*, pages 3923–3927.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.
- Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. *Advances in neural information processing systems*, 34:14681–14693.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.
- Yi-Kun Tang, Heyan Huang, Xuwen Shi, and Xian-Ling Mao. 2023. Neural variational gaussian mixture topic model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–18.
- Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.
- Gary Ulrich. 1984. Computer generation of distributions on the m-sphere. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 33(2):158–163.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.
- Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

- Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Huakui Zhang, Yi Cai, Haopeng Ren, and Qing Li. 2022. Multimodal topic modeling by exploring characteristics of short text social media. *IEEE Transactions on Multimedia*, 25:2430–2445.
- Bingshan Zhu, Yi Cai, and Jiexin Wang. 2024. Graph-based multimodal topic modeling with word relations and object relations. *IEEE Transactions on Multimedia*, 26:8210–8225.
- Elaine Zosa and Lidia Pivovarov. 2022. Multilingual and multimodal topic modelling with pretrained embeddings. In *Proceedings of the 29th international conference on computational linguistics*, pages 4037–4048.

A Reparameterized vMF Sampling

For the hyperspherical latent variable, gradients are propagated through the vMF posterior by using a reparameterized sampler. Given $q(\mathbf{z}_i | \mathbf{x}_i) = \text{vMF}(\boldsymbol{\mu}_i, \kappa_i)$, the sampler first draws a radial variable r_i determined by κ_i :

For $K = 3$, we draw $u_i \sim \text{Uniform}(0, 1)$ and this radial variable can be written as

$$r_i = 1 + \frac{1}{\kappa_i} \log \left(u_i + (1 - u_i) e^{-2\kappa_i} \right) \quad (20)$$

For higher-dimensional hyperspheres, rejection sampling is used for the radial component (Ulrich, 1984). The procedure is shown in Algorithm 1.

Algorithm 1: Acceptance-Rejection Sampling

Input: dimension K , concentration κ

Output: axial component r

```

1 begin
2   Initialize auxiliary parameters;
3    $b \leftarrow \frac{-2\kappa + \sqrt{4\kappa^2 + (K-1)^2}}{K-1}$ ;
4    $a \leftarrow \frac{(K-1) + 2\kappa + \sqrt{4\kappa^2 + (K-1)^2}}{4}$ ;
5    $d \leftarrow \frac{4ab}{1+b} - (K-1) \ln(K-1)$ ;
6   repeat
7     Sample  $\varepsilon \sim \text{Beta}(\frac{K-1}{2}, \frac{K-1}{2})$ ;
8      $r \leftarrow \frac{1 - (1+b)\varepsilon}{1 - (1-b)\varepsilon}$ ;
9      $t \leftarrow \frac{2ab}{1 - (1-b)\varepsilon}$ ;
10    Sample  $u \sim \text{Uniform}(0, 1)$ ;
11    until  $(K-1) \ln t - t + d \geq \ln u$ ;
12    return  $r$ 
13 end
```

A unit vector $\mathbf{s}_i \in \mathbb{R}^{K-1}$ is then sampled to parameterize the angular component orthogonal to the first coordinate before alignment with the posterior mean direction:

$$\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K-1}), \quad \mathbf{s}_i \leftarrow \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|}. \quad (21)$$

The final sample is obtained by first constructing an intermediate vector $\mathbf{z}'_i \in \mathbb{R}^K$ with first coordinate r_i and remaining coordinates $\sqrt{1 - r_i^2} \mathbf{s}_i$, i.e.,

$$\mathbf{z}'_i = \begin{bmatrix} r_i \\ \sqrt{1 - r_i^2} \mathbf{s}_i \end{bmatrix} \in \mathbb{R}^K, \quad (22)$$

and then aligned with the target mean direction $\boldsymbol{\mu}_i$ applying a Householder transformation $h(\cdot)$:

$$\mathbf{z}_i = h(\mathbf{z}'_i, \boldsymbol{\mu}_i) = (\mathbf{I}_K - 2\mathbf{a}_i \mathbf{a}_i^\top) \mathbf{z}'_i, \quad (23)$$

where $\mathbf{a}_i = \frac{\mathbf{e}_1 - \boldsymbol{\mu}_i}{\|\mathbf{e}_1 - \boldsymbol{\mu}_i\|}$.

B KL Divergence on the Hypersphere

The KL regularization term encourages the vMF posterior to stay close to the hyperspherical uniform prior, which corresponds to $\kappa = 0$ and helps prevent overfitting. For a corpus of N documents, the KL term is computed as

$$\mathcal{L}_{\text{KL}} = \frac{1}{N} \sum_{i=1}^N \left[\kappa_i \frac{I_{K/2}(\kappa_i)}{I_{K/2-1}(\kappa_i)} + \log C_K(\kappa_i) - \log \left(\frac{\Gamma(K/2)}{2\pi^{K/2}} \right) \right], \quad (24)$$

where $\Gamma(\cdot)$ is the Gamma function and $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind, defined as:

$$I_\nu(\kappa) = \frac{1}{\pi} \int_0^\pi e^{\kappa \cos(t)} \cos(\nu t) dt. \quad (25)$$

C Implementation Details

We conduct all experiments in PyTorch 2.9 and CUDA 13.0 with a single NVIDIA RTX 5090 GPU, using a batch size of 256 and random seeds {42, 142, 420}, reporting the average results over the three runs.

For all baselines, we use the hyperparameter settings recommended in the original papers when available, and otherwise adopt the released default configurations.

For our method, we construct an image-based neighbor graph with 10 nearest neighbors and keep the top $K_c = 75$ completion words for each document. The graph-completion encoder is trained for 30 epochs with Adam and learning rate 1×10^{-3} . Its loss weights are set to 1.0 for paired contrastive alignment, text reconstruction and image reconstruction, and 2.0 for lexical completion prediction.

The posterior and topic-word concentrations κ and κ_k^t are initialized to 50.0 and learned during training. The temperature scaling factor τ is set to 10. Topic learning is trained for 50 epochs with Adam and learning rate 2×10^{-3} . The weights for observed-BoW reconstruction, text/image reconstruction, and KL regularization are all set to 1.0, and the completion-aware BoW weight is $\lambda_{\text{cbow}} = 0.5$.