Neural Topic Modeling via Contextual and Graph Information Fusion

Jiyuan Liu[1], Jiaxing Yan[1], Chunjiang Zhu[2], Xingyu Liu[1], Qing Li[3], Yanghui Rao[1*]
[1]School of Computer Science and Engineering, Guangdong Key Laboratory of
Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China
{liujy563,yanjx6,liuxy356}@mail2.sysu.edu.cn, czhu@odu.edu,
csqli@comp.polyu.edu.hk,raoyangh@mail.sysu.edu.cn

Abstract

Topic modeling is a powerful unsupervised tool for knowledge discovery. However, existing work struggles with generating limited-quality topics that are uninformative and incoherent, which hinders interpretable insights from managing textual data. In this paper, we improve the original variational autoencoder framework by incorporating contextual and graph information to address the above issues. First, the encoder utilizes topic fusion techniques to combine contextual and bag-of-words information well, and meanwhile exploits the constraints of topic alignment and topic sharpening to generate informative topics. Second, we develop a simple word co-occurrence graph information fusion strategy that efficiently increases topic coherence. On three benchmark datasets, our new framework generates more coherent and diverse topics compared to various baselines, and achieves strong performance on both automatic and manual evaluations. We make our code available for reproduction[1].

1 Introduction

Topic models, which can automatically discover coherent and meaningful topics from text corpora, have been widely used for text data analysis (Rubin et al., 2012; Wang et al., 2018; Jelodar et al., 2020) and knowledge discovery (Wang et al., 2022b; Jelodar et al., 2019). In these methods, each topic is interpreted as a set of related words representing a semantic concept.

Current topic models can be roughly classified into two lines. The first category is methods based on probabilistic graphical models (Blei et al., 2003) or matrix factorization (Kim et al., 2015; Shi et al., 2018). They infer parameters through Gibbs sampling, variational inference, or multiplicative update, which requires high computational costs or complex derivation (Chen et al., 2021, 2023). The second category is neural topic models, including GSM (Miao et al., 2017a), ProdLDA (Srivastava and Sutton, 2017), ETM (Dieng et al., 2020), and so forth (Zhao et al., 2021; Wang et al., 2022a; Wu et al., 2023). These methods generally adopt the Variational Autoencoder (VAE) or Optimal Transport (OT), utilizing back-propagation for high computational efficiency (Wu et al., 2024a).

Model	Label	NPMI	Topic words
FASTopic	religion	0.154	Topic#1 douglas, johnson, influence, link, lord
	baseball	0.084	Topic#2 playoff, local, seattle, open, situation
CGTM	religion	0.368	Topic#1 god faith religion love belief
	baseball	0.380	Topic#2 pitching hitter pitch season players

Table 1: Top 5 related words and NPMI score of the discovered topics by FASTopic (Wu et al., 2024b) and our CGTM on 20News. For clarity, we manually select the topics most relevant to the labels.

However, the current neural topic models struggle with producing limited-quality topics (Hoyle et al., 2022) due to two issues: (1) *uninformative*: topics should be generated to express **core information** within the corpus. As exemplified in Topic#1 of Table 1, the existing FASTopic (Wu et al., 2024b) excessively focuses on fringe topic words like ‘douglas’ and ‘johnson’. (2) *incoherent*: every topic needs to have **internal consistency**. As exemplified in Topic#2 of Table 1, when focusing on baseball-related topic, inconsistent words such as ‘local’ and ‘situation’ should be excluded.

First, in terms of capturing **core information**, some works on representation learning (Federici et al., 2020; Tsai et al., 2021) consider **invariant information shared under different representations as core information that should be retained**. Meanwhile, there are two common representations for documents: one is Bag-of-Words (BoW), especially TF-IDF that recognizes in-domain keywords (Chen and Su, 2023; Haley, 2020). The other uses Pre-trained Language Models (PLMs), espcially

[*]Corresponding author.
[1]https://github.com/Liujyuan/CGTM

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13247–13263
November 4-9, 2025 ©2025 Association for Computational Linguistics

BERT to introduce contextual information (Bianchi et al., 2021). Simply concatenating or adding them together cannot bring much improvement due to their essential differences (Bianchi et al., 2021; Li et al., 2024). Hence how to leverage their collective strengths is still under-explored. Thus, we propose to **capture informative topics by effectively fusing TF-IDF and BERT representations**.

In addition, word relationships can be effective in aiding **consistency within a topic** (Adhya and Sanyal, 2024), so it is becoming a popular practice to exploit the word relationship graph in the topic model. Some Graph Neural Network (GNN)-based works (Zhu et al., 2018; Shen et al., 2021; Adhya and Sanyal, 2024) construct word co-occurrence graphs within documents or sliding windows. However, these studies fail to harness the **overall word co-occurrence at the corpus level** in a well-designed manner that facilitates coherent topic mining.

Based on the aforementioned analysis, we propose solutions aimed at addressing the problems of topics being uninformative and incoherent. As an answer to issue (1), we develop a topic-wise fusion method to merge the TF-IDF and BERT representations. **To capture the informative topic**, we add a topic alignment constraint, which aligns these two sources of representations at the topic level before fusion. Additionally, we adopt topic sharpening, a self-training objective aiming at promoting a stronger emphasis on relevant informative topics in the fused document-topic distribution. As for issue (2), we adopt the fine-grained word relation to tune the word embedding space (Han et al., 2024) through the Graph Information Fusion (GIF). And a simple graph decoder is designed to efficiently improve **topic coherence** by only using the reconstructed word relation graph.

In summary, our contributions are as follows:

- We extend the VAE framework to CGTM, which redefines the text generation process that incorporates **C**ontextual information and **G**raph information into **T**opic **M**odeling.
- For contextual information, we employ a topic-wise fusion approach and impose constraints on both topic alignment and topic sharpening for informative topics.
- For graph information, we employ a graph information fusion method and impose a graph decoder for coherent topics.
- Extensive experiments are conducted on three datasets to evaluate our model. The results show that the performance of CGTM is significantly better than state-of-the-art baselines.

2 Related Works

2.1 Neural Topic Models incorporating Contextual Information

There are three main approaches to integrate contextual information in neural topic models: The first is clustering-based model. These models embed documents into dense vectors using PLMs and then obtain the topic words through various clustering methods. Several approaches, such as BERTopic (Grootendorst, 2022), generate topic representations with the class-based TF-IDF procedure. CETopic (Zhang et al., 2022) further explores different topic vocabulary selection strategies. The second category is data augmentation-based model. This method utilizes PLMs for embedding documents functioning as a means of enhancing the input data or serving as reconstruction targets. BAT (Hoyle et al., 2020) applies knowledge distillation in pre-trained transformers to improve any base neural topic models. CombinedTM (Bianchi et al., 2021) simply concatenates raw BoW and Sentence-BERT (SBERT) embeddings as data augmentation. The third category is embedding-based methods. FastTopic (Wu et al., 2024b) models the semantic relations among document embeddings from BERT and learnable topic and word embeddings using optimal transport. CWTM (Fang et al., 2024) also tries to combine contextual word embeddings.

However, despite the emphasis placed by all aforementioned research on the importance of integrating contextual information into topic modeling, these endeavors frequently fall short, either by isolating the utilization of PLMs or by neglecting the synthesis of TF-IDF and BERT representations. Consequently, they ultimately fail to accurately capture informative topics.

2.2 Neural Topic Models incorporating Graph Information

Due to the rapid development of graph learning (Wu et al., 2021), some of the previous works (Yang et al., 2020; Xie et al., 2021) used GNNs to incorporate graph information into neural topic models. The graph biterm topic model (Zhu et al., 2018), an extension of the biterm topic model (Cheng et al., 2014), represents word co-occurrence information as a graph, where nodes represent words and weighted edges reflect the frequency of corresponding biterms. The graph topic model (Zhou et al., 2020b) constructs document graphs based on TF-IDF scores, capturing relationships with graph convolutions. The Graph Neural Topic Model (GNTM) (Shen et al., 2021) utilizes a directed graph between word nodes to integrate semantic information from documents. The latest work in this area is GINopic (Adhya and Sanyal, 2024) that leverages graph isomorphism networks to capture the correlations between words in each document.

However, these studies ignore word co-occurring patterns in the corpus, which require well-designed strategies to enhance topic coherence. Meanwhile, such GNN-based models can be time-consuming in both construction and training processes (Adhya and Sanyal, 2024).

3 Method

We present the text generation process in Fig. 1 and propose a novel framework for neural topic modeling, as demonstrated by Fig. 2.

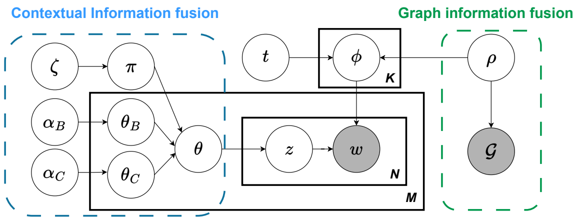


Figure 1: The document generation process of CGTM, where N is the number of words in M documents.

3.1 The Generative Process

As shown in Fig. 1, we describe the following generative process for documents.

1. For each document d in the corpus:

a. Draw the weight $\boldsymbol{\pi} \sim \text{SymDir}(\zeta)$.

b. Draw two topic proportions from two prior distributions $\boldsymbol{\theta}_B \sim \alpha_B, \boldsymbol{\theta}_C \sim \alpha_C$.

c. Draw topic proportion $\boldsymbol{\theta} \sim \varphi\,(\boldsymbol{\theta}_B, \boldsymbol{\theta}_C, \boldsymbol{\pi})$.

2. For each word w in the document:

a. Draw topic assignment $z \sim \text{Cat}(\boldsymbol{\theta})$.

b. Draw the word $w \sim \phi_z$, where topic-word distributions $\phi = \text{Softmax}(t^\top \rho)$.

3. For all words in the corpus, obtain the word relation graph by: $\mathcal{G} \sim \text{Softmax}(\rho^\top \rho)$.

The 2-dimensional vector $\boldsymbol{\pi}$ follows a symmetric Dirichlet distribution with hyper-parameter ζ. α_B and α_C are the piroris of BoW and contextual information, which are $\mathcal{N}\,(0, I)$. $\boldsymbol{\theta} \sim \varphi\,(\boldsymbol{\theta}_B, \boldsymbol{\theta}_C, \boldsymbol{\pi})$ denotes the Gaussian mixture distribution. Further, $\boldsymbol{\theta}$ is the document-topic distribution, t and ρ are the topic and word embeddings, z is the assigned topic, ϕ is the topic-word distribution and $\mathcal{G}$ is the word relation graph.

3.2 Encoder with Contextual Information

Here, we first use the Siamese encoder (Chopra et al., 2005; Yang et al., 2019) to embed TF-IDF and SBERT representations, and then introduce our topic-wise Fusion to obtain the document-topic distribution. Finally, in order to focus on the capture of informative topics, we develop two novel constraints in the encoder.

3.2.1 Siamese Encoder

According to a previous work (Gupta et al., 2023), the siamese encoder has the effect of invariant information learning as well as simplifying the structure, so we adopt the siamese encoder to capture the core information in the TF-IDF and SBERT representations. Here are the details of the siamese encoder.

Given a corpus, each document d is represented by TF-IDF and SBERT, which can be written as:

$$\boldsymbol{x}_B = \text{TF-IDF}(d), \boldsymbol{x}_c = W_C(\text{SBERT}(d)), \quad (1)$$

where $W_C \in \mathbb{R}^{E \times V}$ is a parameter matrix to project the E-dimensional representations through a hidden layer with the same dimensionality as the vocabulary size V.

We follow the framework of VAE (Kingma and Welling, 2014) to infer the topic distribution of document d. In particular, the variational distribution $q(\boldsymbol{\theta}_i, \phi|\boldsymbol{x}_i)$ is an isotropic Gaussian with mean $\boldsymbol{\mu}_i$ and the diagonal of the covariance matrix $\boldsymbol{\sigma}_i^2$ (Miao et al., 2017b), where $i = B, C$. These two parameters are obtained as follows:

$$\boldsymbol{\mu}_i = f_\mu(\boldsymbol{x_i}), \boldsymbol{\sigma}_i^2 = f_\sigma(\boldsymbol{x_i}), i = B, C. \quad (2)$$

In the above, f_μ and f_σ stand for shared multi-layer neural networks for $\boldsymbol{x}_B$ and $\boldsymbol{x}_C$.

By applying the re-parameterization trick for estimation (Kingma and Welling, 2014), we sample

$$\boldsymbol{h}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \boldsymbol{\eta}, i = B, C, \quad (3)$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I})$ denotes an auxiliary noise variable. Then, we obtain the document-topic representations by a softmax function as follows:

$$\boldsymbol{\theta}_i = \text{Softmax}(\boldsymbol{h}_i), i = B, C, \quad (4)$$

where $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C \in \mathbb{R}^K$ denote the document representation over K topics from BoW and contextual information, respectively.

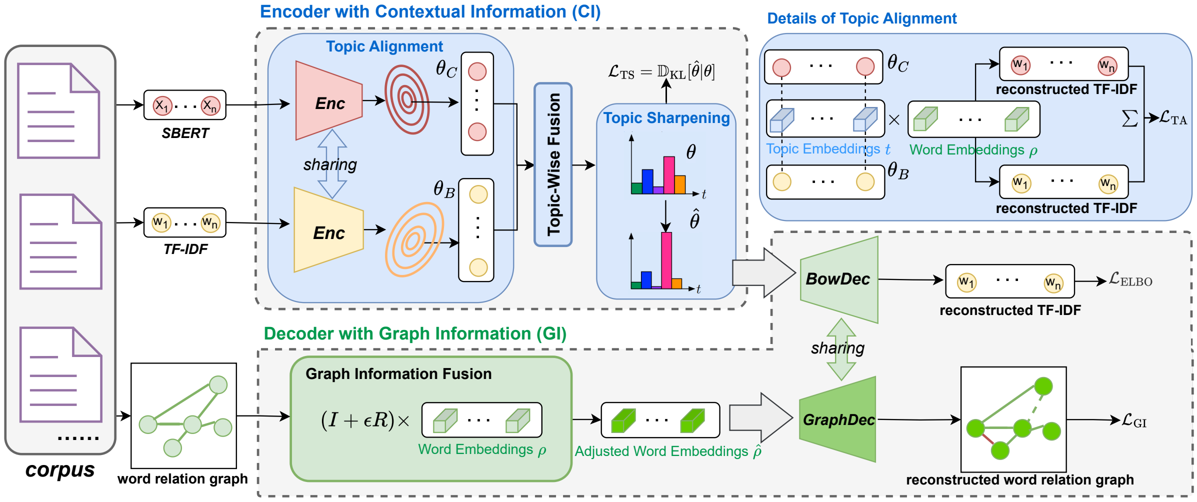

Encoder with Contextual Information (CI)
Topic Alignment
Enc
sharing
Enc
θ_C
θ_B
SBERT
TF-IDF
corpus
word relation graph
Topic-Wise Fusion
$\mathcal{L}_{\text{TS}} = \mathbb{D}_{\text{KL}}[\hat{\theta}|\theta]$
Topic Sharpening
θ
$\hat{\theta}$
Details of Topic Alignment
θ_C
reconstructed TF-IDF
Topic Embeddings t
Word Embeddings ρ
θ_B
reconstructed TF-IDF
$\mathcal{L}_{\text{TA}}$
BowDec
reconstructed TF-IDF
$\mathcal{L}_{\text{ELBO}}$
sharing
Decoder with Graph Information (GI)
Graph Information Fusion
$(I + \epsilon R) \times$
Word Embeddings ρ
Adjusted Word Embeddings $\hat{\rho}$
GraphDec
reconstructed word relation graph
$\mathcal{L}_{\text{GI}}$

Figure 2: The architecture of the proposed CGTM model.

3.2.2 Topic-wise Fusion

Here, we adopt a topic-wise fusion approach to exploit the interconnections between the two document-topic representations within the framework of the Gaussian Mixture Model (GMM), which can be expressed as follows:

$$\boldsymbol{\theta} = \text{Softmax}(Relu(W_f[\tilde{\boldsymbol{\pi}}_B\boldsymbol{\theta}_B; \tilde{\boldsymbol{\pi}}_C\boldsymbol{\theta}_C] + \boldsymbol{b})), \tag{5}$$

where $W_f \in \mathbb{R}^{2K \times K}$ is the weight matrix, $\boldsymbol{b}$ is a learnable bias vector, $[\tilde{\boldsymbol{\pi}}_B, \tilde{\boldsymbol{\pi}}_C]$ is the posterior of the 2-dimensional prior weight $\boldsymbol{\pi}$, and $[\cdot;\cdot]$ represents the concatenation operation.

3.2.3 Topic-level Constraints

As mentioned in a previous work (Li et al., 2024), BERT and TF-IDF representations can be seen as two complementary sources of information for documents. By performing topic-level constraints during the fusion process, we hope to make the model focus on the informative topics.

Therefore, we propose two constraints: **Topic Alignment (TA)** and **Topic Sharpening (TS)**: TA aligns the complex semantics of BERT with the high-frequency keywords of TF-IDF to capture the informative topics. TS makes the informative topics obtained from the fusion step more prominent.

Topic Alignment To align the document-topic representations $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C$ at the topic level, we introduce a cross-entropy term. This term measures the discrepancy between the empirical distribution and the approximate posterior distribution of $\boldsymbol{x_B}$ facilitated through the use of topic embeddings t and word embeddings ρ.

Thus, we define the TA constraint as follows:

$$\begin{aligned}\mathcal{L}_{\text{TA}} &= \sum_{i=B,C} \mathbb{E}_{q(\boldsymbol{\theta}_i,\phi|\boldsymbol{x_i})}[\ln p\left(\boldsymbol{x}_B \mid \boldsymbol{\theta}_i, \phi\right)] \\ &= \sum_{i=B,C} \boldsymbol{x}_B \ln \hat{\boldsymbol{x}}_i,\end{aligned} \tag{6}$$

where $\hat{\boldsymbol{x}}_i = \boldsymbol{\theta}_i \times \text{Softmax}(t^\top \rho)$.

By reconstructing the original TF-IDF representation[2], we ensure that each dimension of $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C$ corresponds to the same topic, **thereby aligning the two types of information at the topic level to capture the informative topics before fusion.**

Topic Sharpening Inspired by DEC (Xie et al., 2016), we employ a self-training objective to utilize representative topics (i.e., topics with high probabilities for each document) as soft labels to sharpen the document-topic distribution.

Particularly, after performing topic-wise fusion, we input the document-topic distribution $\boldsymbol{\theta}$ into a self-training objective, which aims to minimize the following function:

$$\mathcal{L}_{\text{TS}} = \mathbb{D}_{\text{KL}}[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}] = \sum_{z=1}^{K} \hat{\boldsymbol{\theta}}_z \ln \frac{\hat{\boldsymbol{\theta}}_z}{\boldsymbol{\theta}_z}, \tag{7}$$

where $\boldsymbol{\theta}$ is used as the soft topic assignment distribution for each document and $\boldsymbol{\theta}_z$ is the probability that the document belongs to topic z. K is the number of topics. $\hat{\boldsymbol{\theta}}_z$ in Eq. (7) is the target distribution defined as follows:

$$\hat{\boldsymbol{\theta}}_z = \frac{{\boldsymbol{\theta}_z}^2 / f_z}{\sum_{z'=1}^{K} {\boldsymbol{\theta}_{z'}}^2 / f_{z'}}, \tag{8}$$

[2] According to Eq. (1), we introduce W_C to transform the dimension of $\boldsymbol{x}_C$. To avoid a trivial solution (Salehi et al., 2024), we do not reconstruct $\boldsymbol{x}_C$ to get the TA constraint.

where $f_z = \sum_{i=1}^{\mathbf{M}} \theta_z^i$ is the sum of soft topic assignments for all M documents on topic z.

Generally, the target distribution $\hat{\boldsymbol{\theta}}$ raises $\boldsymbol{\theta}$ to the second power to get a sharper document-topic distribution. By minimizing the KL divergence between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, the encoder **focuses more on informative topics after fusion.**

3.3 Decoder with Graph Information

In this section, we design two decoders, BowDecoder for reconstructing the information of each document and GraphDecoder for reconstructing the global word relation graph information.

3.3.1 BowDecoder

In BowDecoder, we first estimate topic-word distributions ϕ by

$$\phi = p\left(\rho | t\right) = \text{Softmax}(t^\top \rho). \quad (9)$$

Then we obtain the reconstructed $\hat{\boldsymbol{x}}$ as:

$$\hat{\boldsymbol{x}} = \boldsymbol{\theta} \times \phi. \quad (10)$$

3.3.2 GraphDecoder

Here, we firstly describe the construction of the global word relation graph, subsequently introduce the graph information fusion method, and ultimately present the graph reconstruction process.

Word Relation Graph Construction A pair of words w_i and w_j is considered co-occurring if their TF-IDF values are simultaneously greater than ξ in the same document, and the total number of co-occurrences for w_i and w_j in the entire corpus is denoted as $A_{i,j}$. We build a weighted undirected graph $\mathcal{G}$ as the normalized word relation matrix $R = (R_{i,j})_{1 \leq i,j \leq V}$, where each elements $R_{i,j}$ is determined as follows:

$$R_{i,j} = A_{i,j} / (D_i \times D_j). \quad (11)$$

In the above, D_i denotes the degree of node i in the word co-occurrence matrix A.

Graph Information Fusion Inspired by LM-Steers (Han et al., 2024), we propose a Graph Information Fusion method (GIF), which takes the word relation graph to tune the original word embeddings as follows:

$$\hat{\rho} = GIF(\rho, R) = \rho + \epsilon R \rho = (I + \epsilon R)\rho, \quad (12)$$

where ϵ is a hyper-parameter related to the corpus. $\hat{\rho}$ is the adjusted word embeddings.

Reconstruction of Graph The posterior distribution of graph $\mathcal{G}$ is estimated by:

$$q(\hat{\mathcal{G}}|\hat{\rho}) = \text{Softmax}(k(\hat{\rho}, \hat{\rho})), \quad (13)$$

where $\hat{\rho} = \text{GIF}(\rho, R)$ represents the use of GIF to incorporate graph information into the original word embeddings. Here, we choose the consine similarity to define our kernel function: $k(\hat{\rho}, \hat{\rho}) = \frac{\hat{\rho}\hat{\rho}^T}{||\hat{\rho}||||\hat{\rho}||}$, and Softmax function is used for normalization. According to the generative process in Section 3.1, we define the prior distribution of graph $\mathcal{G}$ as follows:

$$p(\mathcal{G}) = \text{Softmax}(R). \quad (14)$$

Then, we use the KL divergence to make the posterior distribution of a word relation graph converge to a prior graph as follows:

$$\mathcal{L}_{\text{GI}} = \mathbb{D}_{\text{KL}}[q(\hat{\mathcal{G}}|\hat{\rho}) \parallel p(\mathcal{G})]. \quad (15)$$

By continuously reconstructing the word relation graph, we can fuse fine-grained word relations in the graph decoder practically.

3.4 Overall Loss Function

Our model is derived from the VAE framework, thus we adopt $\mathcal{L}_{\text{ELBO}}$ to maximize the ELBO. Here maximizing ELBO is equivalent to maximizing a variational lower bound on ELBO[3] as follows:

$$\begin{aligned} &\mathcal{L}_{\text{ELBO}} \geq \hat{\mathcal{L}}_{\text{ELBO}} \\ &= \boldsymbol{x_B} \ln(\hat{\boldsymbol{x}}) - \frac{1}{2} \sum_{i=B,C} \left[\boldsymbol{\mu}_i^2 + \boldsymbol{\sigma}_i^2 - \ln\left(\boldsymbol{\sigma}_i^2\right)\right]. \end{aligned} \quad (16)$$

The final loss function, which integrates the constraints of contextual and graph information with ELBO, is presented below for joint training:

$$\mathcal{L} = -\hat{\mathcal{L}}_{\text{ELBO}} + \gamma_C \mathcal{L}_{\text{CI}} + \gamma_G \mathcal{L}_{\text{GI}}, \quad (17)$$

where $\mathcal{L}_{\text{CI}} = \mathcal{L}_{\text{TA}} + \lambda \mathcal{L}_{\text{TS}}$. Besides, λ, γ_C and γ_G are the hyper-parameters that respectively control the weights of TA constraints, contextual and graph information. The training procedure of our CGTM is described in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Datasets: Our experiments are conducted on three widely-used benchmark text datasets, including

[3] Details of the derivation are given in Appendix A.

Algorithm 1: The training procedure of CGTM

```
Input:
    The word embedding $\rho$ from a pre-trained model, #topics $K$, #mini-iterations $I$.
Output: Topic-word distribution $\phi$,
    Document-topic distribution $\boldsymbol{\theta}$.
Construct the word co-occurrence graph $\mathcal{G}$ and compute $R$ as described in Section 3.3.2.
while not converge do
    for $i$ in $1 : I$ do
        Get $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C$ by Siamese Encoder.
        Fuse $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C$ to $\boldsymbol{\theta}$ by Eq. (5).
        Compute $\mathcal{L}_{\text{TA}}$ and $\mathcal{L}_{\text{TS}}$ by Eqs. (6) and (7), respectively.
        Get reconstructed $\boldsymbol{x_B}$ by BowDecoder.
        Get reconstructed $\hat{R}$ by GraphDecoder.
        Update parameters with gradients of $\mathcal{L}$ by Eq. (17).
    end for
end while
```

20News (Miao et al., 2017a), New York Times (NYT) (Meng et al., 2018), and AGnews (Zhang et al., 2015). All datasets have been processed to remove stop words and to filter low frequency words by following (Chen et al., 2023). The statistics of datasets are shown in Table 2, where "#Train" and "#Test" denote the number of documents in training and testing sets, "Vocab" is the vocabulary size, "Avg Len" is the averaged number of words in a document (i.e., document length), and "#Labels" denotes the number of labels.

Table 2: Basic dataset statistics.

Dataset	#Train	#Test	Vocab	Avg Len	#Labels
20News	11,314	7,531	3,997	64.8	20
NYT	7,456	5,233	8,174	272.2	29
AGnews	59,000	3,800	4,422	19.1	4

It is worth emphasizing that these datasets differ in terms of the vocabulary size, the averaged document length, the number of documents and labels. Specifically, AGnews is often used as a short text dataset (Wu et al., 2020), which makes topic modeling difficult due to the sparsity of contextual information (Yan et al., 2013).

Baseline models: For completeness, we compared our CGTM with four groups of existing topic models. Traditional topic models include: 1) **LDA** (Blei et al., 2003) 2) **ProdLDA** (Srivastava and Sutton, 2017). Contextual topic models include: 3) **CombinedTM** (Bianchi et al., 2021), 4) **BERTopic** (Grootendorst, 2022), 5) **CWTM** (Fang et al., 2024), 6) **FASTopic** (Wu et al., 2024b). Graph-based topic models include: 7) **GNTM** (Shen et al., 2021), 8) **GINopic** (Adhya and Sanyal, 2024). VAE-based topic models include: 9) **ETM** (Dieng et al., 2020), 10) **ECRTM** (Wu et al., 2023). The training details are shown in Appendix B.

4.2 Topic-Word Distribution Quality

Interpretability of Topics

We use the widely adopted Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) to evaluate topic interpretability. For completeness, we also include the Coherence Value (CV) (Röder et al., 2015) as a metric.

As shown in Table 3, **our model significantly outperforms other models in terms of topic interpretability**. Models that introduce contextual information from PLM (e.g., FASTopic and CombinedTM) have low topic interpretability due to the lack of topic-level constraints on contextual information, thus preventing access to informative topics. Model that incorporates graphs, such as GINopic (Adhya and Sanyal, 2024) shows relative improvements in coherence, which achieves suboptimal NPMI scores that are about 20% lower than ours. The results confirm the successful capture of informative topics and the improvement in internal topic coherence, thereby effectively guaranteeing topic interpretability (Lau et al., 2014).

Topic Diversity

To measure the semantic redundancy of all topics, we employ the widely-used topic uniqueness (TU) (Nan et al., 2019) for evaluation.

In terms of TU, method that use OT to directly constrain topics and words, such as ECRTM, exhibits relatively high TU score. Nonetheless, **our model maintains optimal TU scores on the majority of datasets as well**. It is worth noting that, as highlighted in previous researches (Wu et al., 2024a; Lu et al., 2024; Liu et al., 2024b), the TU metric does not account for the phenomenon of lexical polysemy. Therefore, a minor degree of repetition is acceptable if it ensures high interpretability.

Topic Quality

Intuitively, better intra-topic consistency may lead to increased redundancy between topics. Conversely, topics with higher TU scores tend to be marginal topics (Wu et al., 2020), typically represented by words with less coherence. Therefore,

Table 3: Topic quality results on three datasets, where the best and the second best results are highlighted in bold and underlined, respectively. * denotes our CGTM improves the best baseline at p-value < 0.05 with paired t-test.

Dataset	Metric	LDA	ProdLDA	ETM	CombinedTM	GNTM	BERTopic	CWTM	ECRTM	GINopic	FASTopic	CGTM
20News	NPMI	0.227	0.236	0.200	0.132	0.161	0.271	0.240	0.149	0.281	0.134	**0.346***
	TU	0.809	0.869	0.266	0.815	0.815	0.783	0.574	**0.956**	0.578	0.899	0.940
	TQ	0.184	0.205	0.053	0.114	0.131	0.212	0.138	0.142	0.162	0.120	**0.325***
	CV	0.510	0.500	0.491	0.302	0.383	0.587	0.553	0.339	0.614	0.309	**0.706***
NYT	NPMI	0.189	0.294	0.133	0.032	0.027	0.287	0.161	0.035	0.279	0.032	**0.323***
	TU	0.636	0.916	0.217	0.919	0.883	0.712	0.638	0.946	0.580	0.950	**0.954***
	TQ	0.120	0.270	0.029	0.030	0.024	0.204	0.103	0.033	0.162	0.031	**0.308***
	CV	0.399	0.584	0.318	0.067	0.054	0.599	0.350	0.073	0.597	0.066	**0.642***
AGnews	NPMI	0.110	0.187	0.064	0.011	0.022	0.167	0.109	0.014	0.217	0.012	**0.264***
	TU	0.936	0.839	0.159	0.794	0.719	0.809	0.790	0.850	0.699	0.872	**0.948***
	TQ	0.103	0.157	0.010	0.009	0.016	0.135	0.086	0.012	0.152	0.011	**0.250***
	CV	0.209	0.359	0.134	0.020	0.041	0.308	0.211	0.025	0.430	0.021	**0.443***

to evaluate the overall quality of topics more comprehensively, we use the topic quality (TQ) metric (Dieng et al., 2020).

In terms of TQ, our model significantly outperforms other models. Specifically, **it outperforms the suboptimal model by 14% on NYT, 53% on 20News, and 59% on AGnews**. These results indicate that our model generates topics that are both highly interpretable and unique. Additionally, the results on AGnews demonstrate that our model achieves superior topic quality without being specifically designed for short texts.

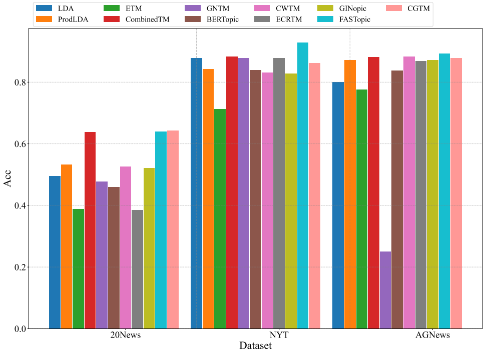


Figure 3: Document classification results on three datasets.

4.3 Document-Topic Distribution Quality

There are some works (Adhya and Sanyal, 2024; Wu et al., 2023) that conduct text classification tasks to evaluate document representational capabilities. Specifically, we use the document-topic distributions obtained from topic models as document features, then train Support Vector Machines (SVMs) to predict the label of each document. We employ Accuracy (Acc) as the evaluation metric, **which reflects the relevance of document-topic distribution and document label.**

As shown in Fig. 3, the performance of CGTM on all datasets is comparable to the optimal model FASTopic at a high level. The good classification results of FASTopic and CombinedTM may be due to their sacrificed topic quality, which retains much of the original PLM embedding information, making them suitable for document classification. These results show that CGTM can **capture high-quality topics while obtaining a reasonable topic distribution of each document**.

Table 4: Result comparison of human evaluation.

	ProdLDA	CombinedTM	FASTopic	GINopic	CGTM
WIS	0.408	0.213	0.216	0.530	**0.750**
TIS	0.594	0.320	0.313	**0.713**	**0.713**

4.4 Human Evaluation

Human evaluation is crucial in developing new topic models, ensuring that the model is consistent with human understanding and expectations (Gao et al., 2024).

In this vein, manual assessments are a necessary complement to automated indicators (Hoyle et al., 2021). Thus we manually estimate both Word Intrusion Score (WIS) and Topic Intrusion Score (TIS) (Chang et al., 2009). The results are shown in Table 4 and the experiments details are shown in Appendix C.

Analysis of WIS

WIS measures the **interpretability of topics**. Similar to NPMI and CV shown in Table 3, our model significantly outperforms the best baseline models by 31% in capturing interpretable topics. At the same time, we argue that the more informative the topic is, the higher the interpretability, as topics that are too marginal are not favorable for human understanding. This highlights that our model mines informative and coherent topics with higher interpretability.

Analysis of TIS

TIS is often used to evaluate the **quality of document-topic distributions** manually. Compared to the metric in Section 4.3, **TIS directly reflects the relevance of the document-topic distribution and the document semantics**, and it is more rational to represent the model's ability in helping humans understand the document at the topic level.

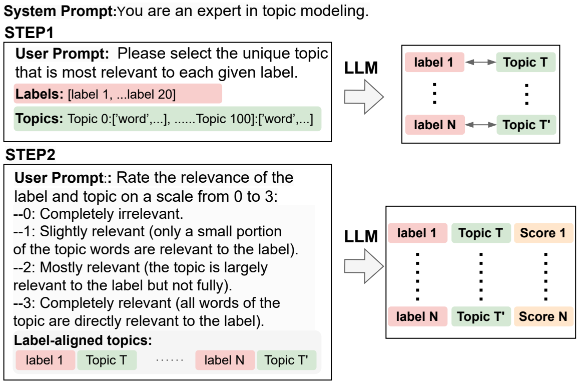
Two points from the results are worth emphasizing: (1) As shown in Table 4, several models that excel in document classification, such as CombinedTM and FASTopic, obtain low TIS results. This may be due to the lack of topic-level constraints on contextual information, which leads to redundant information unrelated to the informative topic and makes the topics poorly relevant to the documents . This indicates that these models fail to enhance human understanding of documents and lose the utmost significance in generating the interpretable document-topic distribution (Hoyle et al., 2021). (2) GINopic's TIS is relatively high. However, due to the high redundancy of topics related with documents (its TU score on 20News is 0.61), it is also difficult to understand and organize documents given many repetitive topics.

In summary, the coherent and informative topics captured by our model are highly interpretable, and the document-topic distribution with a large TIS value also helps us understand documents well.

5 Label Alignment Analysis

In addition to obtaining higher interpretability, the capture of informative topics is reflected in their direct alignment with labels, i.e., labels are the most informative topics (Korenčić et al., 2021). Hence, we employ the labels furnished by the dataset to evaluate the efficacy of detecting potential informative topics across the corpus.

Additionally, with the rising popularity of Large Language Models (LLMs), a recent study (Stammbach et al., 2023) has explored using LLMs to assess topic model performance. Inspired by the above work, we employ LLMs to select the most relevant topics corresponding to the labels and score them for relevance. The TWO-STEP prompt is shown in Fig. 4.



System Prompt: You are an expert in topic modeling.

STEP1

User Prompt: Please select the unique topic that is most relevant to each given label.

Labels: [label 1, ...label 20]

Topics: Topic 0:['word',...], ...Topic 100]:['word',...]

STEP2

User Prompt:: Rate the relevance of the label and topic on a scale from 0 to 3:
--0: Completely irrelevant.
--1: Slightly relevant (only a small portion of the topic words are relevant to the label).
--2: Mostly relevant (the topic is largely relevant to the label but not fully).
--3: Completely relevant (all words of the topic are directly relevant to the label).

Label-aligned topics:

label 1 Topic T ...... label N Topic T'

Figure 4: LLM prompts for label and topic alignment.

The top 5 topic words of the partial label-aligned topics as well as the NPMIs are shown in Table 1, which indicate that the topics are highly correlated with the labels, while high coherence within the topics. More case comparisons as well as discussions are in Section 6.

We further present the NPMI and average relevance scores (ARS) for label-aligned topics in Table 5[4]. We can observe that our model yields topics that are better aligned with the labels and are of much higher coherent than other strong baseline models. The visualization of embedding space, sensitivity analysis and computational analysis are shown in Appendices D, E and F.

Table 5: Result comparison of label-aligned topics.

	ProdLDA	CombinedTM	FASTopic	GINopic	CGTM
NPMI	0.250	0.126	0.129	0.296	**0.381**
ARS	2.632	2.130	1.125	2.675	**2.925**

6 Case Study

Table 6 lists some of these labels and their aligned topics. We report the top 10 NPMI and relevance score for each topic, with unrelated words highlighted in blue. (The RS and Label-aligned topic here are obtained from GPT-4o.)

We can observe that several models show little correlation between aligned topics and labels, e.g., FASTopic and CombinedTM. ProdLDA mixes label-aligned topics with irrelevant topics, which leads to a decrease in NPMI. Meanwhile GINopic corresponds to lower relevance scores on some of the labels, reflecting its weak ability to capture the

[4] To ensure the fairness of the evaluation, we use the following four LLMs with TWO-STEP prompts and calculate their average ARS: Qwen-2.5 (Yang et al., 2024), GLM-4 (GLM et al., 2024), DeepSeek-V3 (Liu et al., 2024a), and GPT-4o (Achiam et al., 2023).

Table 6: Generated topics on 20News with unrelated words marked by blue color.

Models	Label	NPMI	RS	Topic Word Examples
ProdLDA	sci.space	0.377	2	moon, mercury, advertising, probe, wind, lunar, planet, surface, orbit, sky
	rec.sport.baseball	0.265	3	ball, runs, espn, fans, baseball, fan, game, coverage, abc, stadium
	comp.sys.mac.hardware	0.266	3	burst, scsi, mac, quality, printers, printer, laser, postscript, sheet, print
	talk.religion.misc	0.195	2	universe, assertion, physical, god, argument, physically, fish, conclusion, true, proved
	alt.atheism	0.185	2	catholic, pope, think, ms, church, believe, churches, people, traditional, atheists
CombinedTM	sci.space	0.156	1	humans, impossible, dual, month, jones, objects, combined, analog, cambridge, ct
	rec.sport.baseball	0.103	1	eisa, frame, perspective, unlike, van, week, chemistry, son, pushing, batting
	comp.sys.mac.hardware	0.156	1	times, moved, help, caught, association, macs, argument, campus, usage, know
	talk.religion.misc	0.101	1	officer, realized, problem, eternal, oriented, draft, comment, scoring, atheism, structure
	alt.atheism	0.107	1	selling, showing, atheist, single, goal, bi, death, feds, bought, surrounding
FASTopic	sci.space	0.074	1	stayed, rocket, satellites, thoughts, diego, uiuc, knowing, connectors, floppy, beat
	rec.sport.baseball	0.147	2	playoff, local, seattle, open, situation, mechanism, hall, final, observations, altogether
	comp.sys.mac.hardware	0.142	0	ss, wheel, brown, relief, scanner, number, dependent, city, alternatives, numerous
	talk.religion.misc	0.105	2	douglas, johnson, influence, link, lord, kid, cadre, church, muslim, shipped
	alt.atheism	0.087	1	threat, turn, religion, breath, paint, models, narrow, quadra, disclaimer, smoke
GINopic	sci.space	0.271	2	space, cost, moon, money, nasa, idea, launch, real, commercial, billion
	rec.sport.baseball	0.291	2	game, year, games, play, hit, win, team, series, baseball, runs
	comp.sys.mac.hardware	0.152	1	software, buy, mac, help, questions, need, info, appreciated, want, heard
	talk.religion.misc	0.305	3	church, book, john, books, bible, paul, word, catholic, law, jesus
	alt.atheism	0.304	3	believe, true, religion, argument, truth, islam, atheists, atheism, religious, exist
CGTM	sci.space	0.431	3	space, nasa, orbit, shuttle, earth, spacecraft, flight, moon, lunar, solar
	rec.sport.baseball	0.393	3	pitching, hitter, pitch, season, players, league, pitcher, team, player, defensive
	comp.sys.mac.hardware	0.321	3	apple, mac, monitor, nec, quadra, centris, powerbook, lc, sony, macs
	talk.religion.misc	0.411	3	god, faith, religion, love, belief, christianity, beliefs, truth, christians, religions
	alt.atheism	0.434	3	morality, moral, objective , subjective, exist, atheist, god, universe, beings, absolute

informative topic in the corpus. In contrast, **our model is more capable of capturing informative topics while generating highly coherent topics.**

Table 7: Results of the ablation study on 20News.

	NPMI	CV	TU	TQ	Acc	ARS
CGTM	**0.347**	**0.710**	0.937	**0.325**	**0.655**	**2.925**
w/o TA	0.342	0.689	0.705	0.241	0.654	2.850
w/o TS	0.336	0.699	0.933	0.313	0.590	2.888
w/o TWF	0.294	0.654	**0.996**	0.293	0.515	2.765
w/o GD	0.328	0.690	0.913	0.300	0.630	2.888
w/o GIF	0.332	0.698	0.900	0.299	0.637	2.900

7 Ablation Study

We perform ablation experiments on our CGTM model to validate the effectiveness of each component. Table 7 shows the ablation results on 20News, including: 1) without using TA, TS (w/o TA, TS), 2) using the standard GMM to fuse $\boldsymbol{\theta}$ instead of topic-wise fusion (w/o TWF); 3) without using GraphDecoder (w/o GD); 4) without using graph information fusion (w/o GIF); The main observations are as follows:

(1) Removing the TA constraint leads to high topic redundancy caused by the lack of topic alignment between TF-IDF and contextual information, and fails to capture informative topics. And the TS constraint is helpful in improving the overall topic quality as well as downstream performance. (2) Not using TWF results in high TU values, but also in significant degradation of topic interpretability and document representation quality. This indicates that TWF captures the topic interactions between TF-IDF and BERT representations well. (3) The effectiveness of the GraphDecoder is demonstrated by the fact that not using GD leads to a decrease in metrics related to topic coherence. (4) Not using GIF leads to a drop in all metrics, which demonstrates the effectiveness of this simple graph learning approach we used.

8 Conclusion

In this paper, we present a novel framework for neural topic modeling via contextual and graph information fusion. We perform a novel topic-wise fusion method between TF-IDF and pre-trained **contextual information**. Meanwhile, we design two constraints: topic alignment and topic sharpening to ensure that the informative topics are captured and the interpretability of the topic-word distribution is enhanced. The **graph information** is fused through a GIF method and GraphDecoder using the overall word co-occurrence graph to obtain good topic coherence. CGTM generates coherent, distinctive and informative topics, outperforming strong baselines in both topic quality and interpretable document-topic distributions.

Limitations

We view our framework as a preliminary step toward extracting informative and coherent topics, as well as generating interpretable document-topic distributions. Three primary limitations remain for future exploration: (1) While we obtain topics with good interpretability by fusing contextual information through two constraints, directly enhancing topic interpretability is still challenging. Future work could benefit from explicitly leveraging LLMs' strong capabilities of language understanding, such as TopicGPT (Pham et al., 2024), a prompt-based framework that requires the user to provide manually-curated example topics first. In principle this is a weakly-supervised problem, so there is still room to extend our method to such areas. (2) The metric for assessing document-topic distribution's interpretability might have weaknesses: we use topic intrusion without accounting for topic repetition, potentially inflating scores (e.g., in GINopic's results), and it relies on time-intensive manual evaluations that could be accelerated using LLMs carefully. (3) Despite our model's commendable performance and acceptable efficiency, delving into methodologies that could further reduce time complexity remains a promising avenue for future exploration.

Acknowledgements

The work of Yanghui Rao has been supported by the National Natural Science Foundation of China (62372483). The work of Chunjiang Zhu has been supported by Natural Science Foundation CCF-2349369. The work of Qing Li has been supported by the Hong Kong Research Grants Council under the Collaborative Research Fund (project no. C1043-24G) and Theme-based Research Scheme (project no. T43-513/23-N).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Janko Altenschmidt Diogo Almeida, Sam Altman, and Shyamal Anadkat et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Suman Adhya and Debarshi Kumar Sanyal. 2024. Ginopic: Topic modeling with graph isomorphism network. In *NAACL-HLT*, pages 6171–6183.

Junwen Bai, Shufeng Kong, and Carla P Gomes. 2022. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *ICML*, pages 1383–1398.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *ACL*, pages 759–766.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NeurIPS*, pages 288–296.

Hegang Chen, Pengbo Mao, Yuyin Lu, and Yanghui Rao. 2023. Nonlinear structural equation model guided gaussian mixture hierarchical topic modeling. In *ACL*, pages 10377–10390.

Jiayang Chen and Qinliang Su. 2023. Exploiting multiple features for hash codes learning with semantic-alignment-promoting variational auto-encoder. In *NLPCC*, pages 563–575.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021. Tree-structured topic modeling with nonparametric neural variational inference. In *ACL*, pages 2343–2353.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Zheng Fang, Yulan He, and Rob Procter. 2024. CWTM: Leveraging contextualized word embeddings from BERT for neural topic modeling. In *LREC-COLING*, pages 4273–4286.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. In *ICLR*.

Xin Gao, Yang Lin, Ruiqing Li, Yasha Wang, Xu Chu, Xinyu Ma, and Hailong Yu. 2024. Enhancing topic interpretability for neural topic modeling through topic-wise contrastive learning. In *ICDE*, pages 584–597.

Team GLM, Aohan Zengand, Bin Xu, and Bowen Wang et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. 2023. Siamese masked autoencoders. In *NeurIPS*.

Coleman Haley. 2020. This is a BERT. Now there are several of them. Can they generalize to novel words? In *BlackboxNLP@EMNLP*, pages 333–341.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *ACL*, pages 16410–16430.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *NeurIPS*, pages 2018–2033.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving neural topic models using knowledge distillation. In *EMNLP*, pages 1752–1771.

Alexander Miserlis Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *EMNLP Findings*, pages 5321–5344.

Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211.

Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. 2015. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *KDD*, pages 567–576.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Damir Korenčić, Strahil Ristov, Jelena Repar, and Jan Šnajder. 2021. A topic coverage approach to evaluation of topic models. *IEEE Access*, 9:123280–123312.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Zetong Li, Qinliang Su, Shijing Si, and Jianxing Yu. 2024. Leveraging bert and tfidf features for short text clustering via alignment-promoting co-training. In *EMNLP*, pages 14897–14913.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, and Chong Ruan et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Guojun Liu, Yang Liu, Maozu Guo, Peng Li, and Mingyu Li. 2019. Variational inference with gaussian mixture model and householder flow. *Neural Networks*, 109:43–55.

Jiyuan Liu, Hegang Chen, Chunjiang Zhu, and Yanghui Rao. 2024b. Unsupervised hierarchical topic modeling via anchor word clustering and path guidance. In *EMNLP Findings*, pages 7505–7517.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *CIKM*, pages 983–992.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *AAAI*, pages 6826–6833.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017a. Discovering discrete latent topics with neural variational inference. In *ICML*, pages 2410–2419.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017b. Discovering discrete latent topics with neural variational inference. In *ICML*, pages 2410–2419.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *ACL*, pages 6345–6381.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. In *NAACL-HLT*, pages 2956–2984.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408.

Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning*, 88:157–208.

Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. 2024. Sigma: Sinkhorn-guided masked video modeling. In *ECCV*, pages 293–312.

Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. In *NeurIPS*, pages 14681–14693.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *WWW*, pages 1105–1114.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. In *NeurIPS*, pages 16857–16867.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *EMNLP*, pages 9348–9357.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. Octis: Comparing and optimizing topic models is simple! In *EACL*, pages 263–270.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Self-supervised learning from a multi-view perspective. In *ICLR*.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. CluHTM - semantic hierarchical topic modeling based on CluWords. In *ACL*, pages 8138–8150.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022a. Representing mixtures of word embeddings with mixtures of topic embeddings. In *ICLR*.

Weifan Wang, Xiaocheng Cheng, Ziqi Liu, Yu Lin, Yue Shen, Binbin Hu, Zhiqiang Zhang, Xiaodong Zeng, Jun Zhou, Jinjie Gu, and Minnan Luo. 2022b. Intent mining: A social and semantic enhanced topic model for operation-friendly digital marketing. In *ICDE*, pages 3254–3267.

Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2018. Topic compositional neural language model. In *AISTATS*, pages 356–365.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *ICML*, pages 37335–37357.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *EMNLP*, pages 1772–1782.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):1–30.

Xiaobao Wu, Thong Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *NeurIPS*.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487.

Qianqian Xie, Jimin Huang, Pan Du, and Min Peng. 2021. Graph relational topic model with higher-order graph attention auto-encoders. In *ACL Findings*, pages 2604–2613.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *WWW*, pages 1445–1456.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*.

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *WWW*, pages 144–154.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*, pages 649–657.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2022. Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics. In *NAACL-HLT*, pages 3886–3893.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021. Neural topic model via optimal transport. In *ICLR*.

Cangqi Zhou, Hao Ban, Jing Zhang, Qianmu Li, and Yinghua Zhang. 2020a. Gaussian mixture variational autoencoder for semi-supervised topic modeling. *IEEE Access*, 8:106843–106854.

Deyu Zhou, Xuemeng Hu, and Rui Wang. 2020b. Neural topic modeling by incorporating document relationship graph. In *EMNLP*, pages 3790–3796.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *EMNLP*, pages 4663–4672.

A Variational Objective

Typically, the log likelihood of the data $\ln p(\boldsymbol{x})$ is maximized for parameter estimation. However, the marginal probability $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is computational intractable (Zhou et al., 2020a). The framework of VAE aims to deal with the above problem (Kingma and Welling, 2014). As a VAE-based neural topic model, our CGTM can be trained by directly maximizing the following Evidence Lower BOund (ELBO):

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} =& \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{x})}[\ln p(\boldsymbol{x}|\boldsymbol{\theta})] \\ &- \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}|\boldsymbol{x})||(p(\boldsymbol{\theta})],\end{aligned} \tag{18}$$

where the first term is the expected log-likelihood and the second one is the KL divergence from the variational posterior $q(\boldsymbol{\theta}|\boldsymbol{x})$ to the prior $p(\boldsymbol{\theta})$. Based on the GMM assumption mentioned in Section 3, $q(\boldsymbol{\theta}|\boldsymbol{x}) = \varphi(\boldsymbol{\theta}_B, \boldsymbol{\theta}_C, \tilde{\boldsymbol{\pi}})$.

Similar to previous researches (Dilokthanakul et al., 2016; Bai et al., 2022), we set the posterior $\tilde{\boldsymbol{\pi}} = [1/S, ...1/S]$ by following the widely-used GMM hypotheses, where S is the predefined number of components in the mixture. Here we have two text representations $\boldsymbol{x}_B$ and $\boldsymbol{x}_C$ i.e., $S = 2$, and $\tilde{\boldsymbol{\pi}} = [0.5, 0.5]$, thus $q(\boldsymbol{\theta}|\boldsymbol{x}) = \tilde{\boldsymbol{\pi}}_B\boldsymbol{\theta}_B + \tilde{\boldsymbol{\pi}}_C\boldsymbol{\theta}_C$, where $\tilde{\boldsymbol{\pi}}_B = \tilde{\boldsymbol{\pi}}_C = 0.5$.

Since the standard Gaussian distribution can be rewritten as a mixture of multiple standard Gaussian distributions, its weights can be arbitrarily normalized as follows:

$$p(\boldsymbol{\theta}) = \mathcal{N}(0, I) = \sum_{i=B,C} \boldsymbol{\pi}_i \alpha_i, \tag{19}$$

$$\sum_{i=B,C} \boldsymbol{\pi}_i = 1, \tag{20}$$

where the prior weight $\boldsymbol{\pi}$ obeys $\mathrm{SymDir}(\zeta)$.

The KL divergence between two probability mixtures $q = \sum_{i=B,C} \tilde{\boldsymbol{\pi}}_i \tilde{f}_i$ and $p = \sum_{i=B,C} \boldsymbol{\pi}_i f_i$ is upper bounded as follows (Liu et al., 2019):

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}\left[q \parallel p\right] \le& \mathbb{D}_{\mathrm{KL}}\left[\tilde{\boldsymbol{\pi}} \parallel \boldsymbol{\pi}\right] \\ &+ \sum_{i=B,C} \tilde{\boldsymbol{\pi}}_i \mathbb{D}_{\mathrm{KL}}\left[\tilde{f}_i \parallel f_i\right] \\ &= \sum_{i=B,C} \tilde{\boldsymbol{\pi}}_i \left[\ln \frac{\tilde{\boldsymbol{\pi}}_i}{\boldsymbol{\pi}_i} + \mathbb{D}_{\mathrm{KL}}\left[\tilde{f}_i \parallel f_i\right]\right],\end{aligned} \tag{21}$$

where the prior weight $\boldsymbol{\pi}_B = \boldsymbol{\pi}_C = 0.5$ approximately by setting a large value for hyper-parameter ζ, in order to maintain simplicity in the derivation. Then, the above equation can be rewritten as:

$$\mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}|\boldsymbol{x})||p(\boldsymbol{\theta})) \le \frac{1}{2} \sum_{i=B,\,C} \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}_i|\boldsymbol{x})||\alpha_i). \tag{22}$$

Finally, $\mathcal{L}_{\mathrm{ELBO}}$ has the following lower bound:

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} \ge& \hat{\mathcal{L}}_{\mathrm{ELBO}} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{x})}[\ln p(\boldsymbol{x}|\boldsymbol{\theta})] \\ &- \frac{1}{2} \sum_{i=B,\,C} \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}_i|\boldsymbol{x})||\alpha_i).\end{aligned} \tag{23}$$

For the right part above, the first term is the reconstruction of $\boldsymbol{x_B}$ by $\boldsymbol{\theta}$ which can be written as $\boldsymbol{x_B}\ln(\hat{\boldsymbol{x}})$. The second term can be $\frac{1}{2}\sum_{i=B,\,C}\left[\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2 - \ln\left(\boldsymbol{\sigma}_i^2\right) - 1\right]$, when $q(\boldsymbol{\theta}_i|\boldsymbol{x})$=$N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2), i = B, C$, and the constant term -1 in the second term can be ignored.

B Training Details

All baselines are run with public model codes either from their source codes or from OCTIS (Terragni et al., 2021), trained on a workstation equipped with an Nvidia RTX 3090 GPU and a Python environment with 24G memory. To keep simplicity, for the multilayer neural networks f_μ and f_σ in the encoder, we use a fully-connected neural network with $Tanh$ as the activation function. For embedding-based topic models including ETM, ECRTM, CWTM and CGTM, we incorporate pretrained word embeddings (Viegas et al., 2020) into them, and the embedding size of word is 300. For the contextual topic models including CombinedTM, BERTopic and CGTM, we use XLNET (Song et al., 2020) without fine-tuning as SBERT, and the embedding size E of SBERT is 768.

For each model, we save the highest topic quality (TQ) during training and treat it as the best one. We construct a word co-occurrence graph and train our model both on the training set. To ensure fair comparisons, we evaluate different models on the same unseen test data. We report averaged scores of 5 runs to obtain statistically stable results.

For hyper-parameters in CGTM, such as γ_C, γ_G, λ and ϵ, we set them as follows:

- 20News dataset: $\gamma_C = 1$, $\gamma_G = 0.2$, $\lambda = 20$, $\epsilon = 0.004$.
- NYT dataset: $\gamma_C = 10$, $\gamma_G = 0.1$, $\lambda = 20$, $\epsilon = 0.004$.
- AGnews dataset: $\gamma_C = 2$, $\gamma_G = 0.2$, $\lambda = 40$, $\epsilon = 0.2$.

The threshold ξ for constructing the word relation graph is set to 1. The hyper-parameter ζ of the symmetric Dirichelet distribution is set to 10. The number of mini-iterations $I = 500$ and the batch size $B = 512$. We set ϵ to $4e-3$ for most datasets, with the variation in ϵ for the AGnews dataset being attributed to its nature as a short text dataset[5]. More details about ϵ's setup are shown in Appendix E.

C Details of Human Evaluation

Word Intrusion

Based on the research conducted in a pervious work (Hoyle et al., 2021), we adopt a word intrusion task as human evaluation on topic interpretability. This task consists of identifying words that do not belong to a coherent potential category represented by the top words in a topic.

To ensure a comprehensive assessment of the topics, we randomly select 25 of 100 topics and randomly use top 1 word from the remaining topics as random intruder word. Each of these topics is presented with its top 10 words.

Topic Intrusion

We utilize a manual evaluation task of topic intrusion (Chang et al., 2009) to further assess the ability of topic models to represent documents. The topic intrusion task consists of identifying the topics that are least relevant to the document.

To ensure a comprehensive assessment of the quality of the document representations in the corpus, we randomly select one document under each label, based on its top 3 topics, as well as randomly selecting a remaining topic as intruder topics. Each of these topics is presented with its top 5 words.

We recruit eight graduate students majoring in computer science as participants and guide them to accomplish the word intrusion and topic intrusion tasks. The word intrusion scores (WIS) and topic intrusion scores (TIS) ranging from 0 to 1, quantitatively measure the ability of annotators to detect the "intruder" word and topic, respectively.

D Visualization of Embedding Space

The top 5 words from 6 topics generated by CGTM on 20News are visualized in Fig. 5 via UMAP (McInnes et al., 2018). We can observe that the topics are embedded in the middle of related words, expressing certain semantic information. Besides, words under the same topic are closer, while words under different topics are farther apart. Furthermore, related topics are closer in the embedding space, such as Topic: 53 *(space)* and Topic:31 *(science)*, Topic: 25 *(atheism)* and Topic:46 *(religion)*, Topic: 43 *(government)* and Topic: 90 *(president)*. It is also worth noting that words with similar semantics in different topics will approach each other, such as *"scientific"* in Topic: 31 and *"nasa"* in Topic: 53, as well as *"moral"* in Topic: 25 and *"religion"* in Topic: 46.

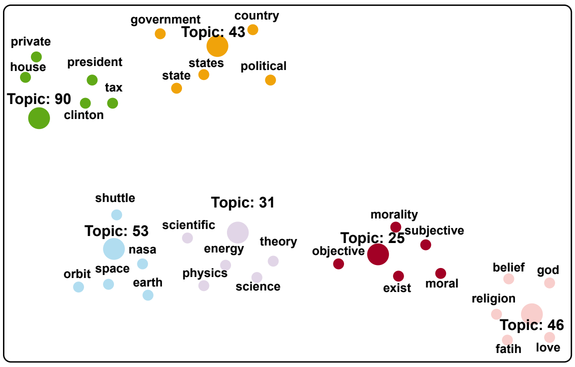
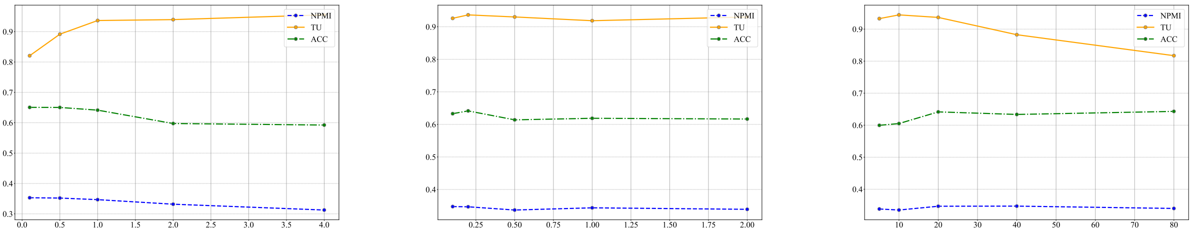


Figure 5: Visualization of word and topic embeddings, where Topic: i denotes the i^{th} topic.

E Sensitivity Analysis

The influence of loss weights The sensitivity analysis is conducted on three hyper-parameters of loss weights. γ_C and γ_G control the weight of contextual and graph information, respectively. λ controls topic sharpening after contextual information is fused. We report the NPMI, TU, and Acc results of our model on 20News in Fig. 6, where γ_c varies between 0.1 and 4, γ_g varies between 0.1 and 2, and λ varies between 5 and 80.

[5] Using $\epsilon R \rho$ in Eq. (12) achieves the equivalent effect as $k\epsilon \cdot k^{-1}R$ when $k \neq 0$, so the absolute value of ϵ itself is only meaningful when also considering the magnitude of R. Since word co-occurrences in the short text are very sparse (Yan et al., 2013), the elements of R are sparse and small, and ϵ should be larger.



(a) γ_c (b) γ_g (c) λ

Figure 6: Performances under different values of hyper-parameters on 20News.

From Fig. 6, we can observe that as γ_c increases, TU score increases, while Acc and NPMI scores decrease. This is because γ_c controls the contextual information constraints, and increasing it causes the model to focus more on informative topics.

For γ_g, the trend is relatively smooth, reflecting the stability of the graph structure for word embedding space constraints.

For the weight λ of the TS constraints in $\mathcal{L}_{\text{TS}}$, NPMI score remains relatively stable, TU score initially rises and then decreases substantially, while Acc score continues to rise. These observations emphasize the significance of a moderate λ value. While an excessively large λ can enhance topic salience, it may lead to increased similarity among topics, and boost redundancy. Therefore, finding an optimal λ that balances salience and uniqueness is crucial for generating high-quality topics.

A clear tradeoff emerges between topic coherence and diversity, which has been acknowledged in previous researches (Wu et al., 2020; Gao et al., 2024). Our experiments also unveil tradeoffs between topic diversity and document classification accuracy. The underlying reason might be that, while maintaining high-quality topics, an excessively high TU score causes the model to focus more on marginal topics, thereby weakening its representational capacity for the document.

Overall, the contrasting trends observed in the metrics depicted in Figs. 6(a) and 6(c), as γ_c and λ increase, respectively, uphold the hypothesis proposed in Section 3.2.3. This hypothesis suggests that the constraints of topic alignment and topic sharpening are mutually reinforcing, ultimately enhancing the core information of topics.

How to set the proper value of ϵ A sensitivity analysis is also conducted on hyper-parameters ϵ, which controls the power of our GIF module.

As mentioned above, when the value of ϵ becomes large, there is a clear trade-off between topic coherence and topic diversity. According to Fig. 7, with the increase of ϵ, more word co-occurrence information is added to our framework. It results in gradual improvements in NPMI and Acc scores, but TU score decreases, indicating the increase in both topic coherence and topic redundancy.

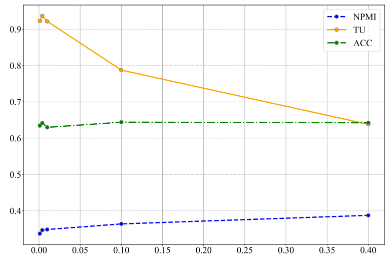


Figure 7: Performances under different values of ϵ on 20News.

In summary, we demonstrate through ϵ's variation that GIF can be used as a controlled and interpretable module to adjust word embeddings through the word relation graph. By tuning the value of ϵ, it is possible to improve topic coherence while maintaining uniqueness among topics.

F Computational Analysis

Theoretically complexity The main time and space requirements in our CGTM can be attributed to the CI and GI modules. For the CI module, the fusion module has a time complexity of $O(B \cdot 2K^2)$, TA has $O(B \cdot 2KVE + B \cdot K)$, and TS has $O(B \cdot K)$. For the GI module, the GraphDecoder has a time complexity of $O(V^2E + V^2)$, and GIF has $O(VE)$. The word co-occurrence graph $\mathcal{G}$ can be pre-constructed, and SBERT data $\boldsymbol{x}_C$ can be pre-generated using PLMs, reducing the computational burden during training. The CI module increases space complexity by $O(M)$ due to $\boldsymbol{x}_C$, and the GI module adds $O(V^2)$ space complexity due to word relation matrix R.

Runtime and Parameter Size The runtime of 500 training epochs and the parameter size of re-

Table 8: Runtime and parameter size on 20News.

	CombinedTM	CWTM	ECRTM	GINopic	FASTopic	CGTM
Runtime	255s	10960s	927s	8501s	64s	1505s
#Params	4.30M	134.89M	2.12M	12.73M	3.05M	3.99M

cent strong models on 20News are shown in Table 8. Among them, FASTopic (Wu et al., 2024b) is relatively fast, but the topic quality and document-topic interpretability are relatively poor according to Tables 3 and 4. From the results in Tables 3, 4 and 8, it can be concluded that our model achieves good performance with acceptable efficiency by fusing contextual and graph information.

G Scalability Analysis

Supplementary Experiment on Large Corpus To further demonstrate the scalability of our model, we here supplement the experiment on the Arxiv dataset (Meng et al., 2019), a set of paper abstracts covering 53 classes from the Arxiv website, which is a much larger corpus. Table 9 provides detailed statistics for this dataset.

Table 9: Arxiv and WoS dataset statistics.

Dataset	#Train	#Test	Vocab	Avg Len	#Labels
Arxiv	131160	92042	11799	57.2	53
WoS	14,095	4,699	8,232	105.1	7

The results are detailed in Table 10. Our model demonstrates an acceptable runtime (about 1 hour) and a manageable number of parameters (11.54M). Furthermore, as shown in Table 11, the performance of our model on Arxiv and WoS is superior to that of other strong baselines. These findings collectively validate the scalability of our model on large corpus.

Table 10: Runtime and parameter size on Arxiv.

	CombinedTM	ProdLDA	GINopic	FASTopic	CGTM
Runtime	2504s	3542s	17002s	2270s	3620s
#Params	12.64M	2.39M	36.27M	8.21M	11.54M

Complexity Comparison with Standard NTMs For the standard NTMs (such as ETM), the time complexity is $\mathcal{O}(B \times (VK + EK + EV))$. In the encoder of our method, the complexity is $\mathcal{O}(B \times (VK + 2K^2 + K))$. Since $K \ll V$, the asymptotic complexity keeps the same.

For the extra graph decoder, while the original word relation graph R has V^2 time and space complexity, we address this limitation via sparse matrix storage and manipulation. On the Arxiv dataset, the above strategy reduces non-zero elements to 5.8% of the original dense matrix.

Table 11: Performance of different models on Arxiv and WoS. * denotes our CGTM improves the best baseline at p-value < 0.05 with paired t-test.

Dataset	Metric	ProdLDA	CombinedTM	FASTopic	GINopic	CGTM
Arxiv	NPMI	0.236	0.018	0.038	0.209	**0.242***
	TU	0.924	0.877	0.964	0.670	**0.978***
	TQ	0.218	0.016	0.037	0.141	**0.237***
	CV	0.513	0.037	0.081	0.490	**0.542***
WoS	NPMI	0.218	0.033	0.034	0.238	**0.248***
	TU	0.883	0.880	0.961	0.505	**0.937***
	TQ	0.193	0.069	0.069	0.120	**0.232***
	CV	0.467	0.069	0.032	0.528	**0.529***

In summary, both the acceptable running cost and theoretical complexity demonstrate the scalability of our CGTM.

H Topic Clustering Quality and Interpretability

There are also many works (Pham et al., 2024; Wu et al., 2024b) adopt Purity, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI) to measure the topic clustering quality of topic model. The Table 12 presents these results of topic Clustering alongside our original metrics, highlighting a critical trade-off between raw clustering performance and topic interpretability, which constitutes a vital insight for the field.

Table 12: Performance of different models on 20News on topic clustering quality and interpretability

Model	NPMI (Topic) Coherence	TIS (Doc.) Interpretability	Purity (Doc.) Clustering	ARI (Doc.) Clustering	NMI (Doc.) Clustering
ProdLDA	0.237	0.408	0.229	0.045	0.287
CombinedTM	0.132	0.213	0.248	0.039	0.329
GINopic	0.284	**0.713**	0.227	0.031	0.293
FASTopic	0.134	0.313	**0.465**	**0.256**	**0.471**
CGTM (Ours)	**0.347**	**0.713**	0.316	0.140	0.329

Corroborating the analysis in Sections 4.3 and 4.4, these results reveal two important conclusions. First, they highlight the steep cost in interpretability for models that pursue high clustering or document classification scores. Second, they demonstrate CGTM's superior ability to balance these competing objectives.

The Price of High Clustering Scores: Models like FASTopic excel in clustering metrics (Purity, ARI, NMI) by retaining a large amount of raw PLM information. However, this comes at a steep cost: their topics are significantly less coherent (low NPMI), and their document representations are poorly aligned with human judgment (low TIS).

This approach sacrifices the core goal of topic modeling: generating meaningful and interpretable document-topic distributions.

CGTM's Superior Balance: In contrast to the above issue, our model, CGTM, strikes a compelling balance. By effectively fusing contextual and graph information and applying our novel topic-level constraints, CGTM produces document representations that are not only highly competitive for clustering (outperforming all baselines except FASTopic) but are also grounded in semantically coherent and highly interpretable topics (as shown by our leading NPMI and TIS scores).

This analysis indicates that CGTM is adept at grouping documents into "truly suitable topics" by producing clusters that are both well-separated in the feature space and defined by semantically coherent concepts for human interpretation.