CEMTM: Contextual Embedding-based Multimodal Topic Modeling

Amirhossein Abaskohi[1], Raymond Li[1], Chuyuan Li[1], Shafiq Joty[2], Giuseppe Carenini[1]

[1] Department of Computer Science The University of British Columbia
V6T 1Z4, Vancouver, BC, Canada
[2] Salesforce AI Research
{aabaskoh, raymondl, carenini}@cs.ubc.ca, chuyuan.li@ubc.ca
sjoty@salesforce.com

Abstract

We introduce **CEMTM**, a context-enhanced multimodal topic model designed to infer coherent and interpretable topic structures from both short and long documents containing text and images. CEMTM builds on fine-tuned large vision language models (LVLMs) to obtain contextualized embeddings, and employs a distributional attention mechanism to weight token-level contributions to topic inference. A reconstruction objective aligns topic-based representations with the document embedding, encouraging semantic consistency across modalities. Unlike existing approaches, CEMTM can process multiple images per document without repeated encoding and maintains interpretability through explicit word-topic and document-topic distributions. Extensive experiments on six multimodal benchmarks show that CEMTM consistently outperforms unimodal and multimodal baselines, achieving a remarkable average LLM score of 2.61 (1-3 scale). Further analysis shows its effectiveness in downstream few-shot retrieval and its ability to capture visually grounded semantics in complex domains such as scientific articles[1].

1 Introduction

Topic modeling aims to uncover the latent thematic structure of a corpus by organizing documents into interpretable clusters of topics. While classical topic models like latent dirichlet allocation (LDA) (Blei et al., 2003) have long been applied to textual corpora, the rapid growth of multimodal content, where images, captions, and structured text co-exist, demands models that can jointly understand and reason over multiple modalities. Traditional multimodal topic models (Feng and Lapata, 2010; Putthividhy et al., 2010) extended LDA to incorporate image features alongside text, but often failed to capture deeper cross-modal interactions.

Recent advances in neural topic modeling (Zhu et al., 2024; Gonzalez-Pizarro and Carenini, 2024a) have addressed some of these limitations by learning shared embeddings across modalities, enabling more coherent and semantically unified topic discovery.

Parallel to these developments, large language models (LLMs) and large vision language models (LVLMs) have shown remarkable capacity to encode rich semantic knowledge from vast and diverse corpora. In text-based topic modeling, LLMs have been used both for generating and assigning topic with zero- and few-shot prompting (Mu et al., 2024; Pham et al., 2024b), significantly improving topic coherence and interpretability. In multimodal settings, early efforts have used prompt-based methods (Prakash et al., 2023). However, while models like TopicGPT produce interpretable outputs through natural language, they lack corpus-level topic distributions and robustness to prompt variation. They also do not model uncertainty or provide consistent global topic structures, limiting their usefulness for exploratory analysis (Hosseiny Marani and Baumer, 2023). A promising direction is to combine the knowledge grounding and modality alignment of LVLMs with the structured modeling of multimodal neural topic models, leveraging LVLMs to enhance semantic understanding without compromising the coherence and stability of topic representations.

To address these limitations, we propose **CEMTM** (Contextual Embedding-based Multimodal Topic Modeling), a novel topic modeling framework that directly leverages the latent representations produced by pretrained LVLMs. Instead of designing complex architectures to align modalities, CEMTM uses the final token embedding from an LVLM as a compact, unified representation of a multimodal document that contains both textual and visual content (Jiang et al., 2025). This approach not only captures deeply aligned cross-

---

[1]Code is publicly available at: https://github.com/AmirAbaskohi/CEMTM.

11675

modal semantics but also simplifies the processing of documents with multiple images. By avoiding the need for separate modality-specific encoders, CEMTM allows the entire document, including all images and the accompanying text, to be encoded holistically, making it well-suited for scalable and coherent multimodal topic modeling. Additionally, inspired by Fang et al. (2024), we incorporate a learnable importance network to estimate the contribution of each textual token and image patch to the document-topic representation. CEMTM achieves strong empirical performance across six benchmark datasets, obtaining an average LLM coherence score of **2.61** out of 3, outperforming a broad range of baselines.

Our contributions are: **(I)** We introduce **CEMTM**, a multimodal topic model that uses pretrained vision-language representations to generate coherent, diverse topics from long multimodal documents; **(II)** We propose a stochastic, distribution-based mechanism to learn token importance, improving semantic alignment and interpretability when combined with fine-tuned LVLM embeddings; **(III)** CEMTM sets a new SOTA, outperforming strong baselines on topic quality and the downstream task of few-shot question-answering (QA), demonstrating the value of topic distributions for retrieval-based tasks.

2 Related Work

Neural Multimodal Topic Modeling Early multimodal topic models extended LDA to handle image and text jointly (Blei and Jordan, 2003), but often treated modalities independently. Neural approaches addressed this by learning shared representations, such as SupDocNADE (Zheng et al., 2014) and graph-based models for short documents (Zhu et al., 2024). Gonzalez-Pizarro and Carenini (2024b) conducted a large-scale comparison of neural multimodal topic models, showing room for improvement in coherence and diversity. Unlike these models, CEMTM leverages pretrained LVLMs and uses their final token embeddings to capture aligned cross-modal semantics, eliminating the need to learn modality alignment during topic representation learning.

Language Models for Topic Modeling Language models have advanced topic modeling through prompting and contextual embeddings. Prompt-based methods like TopicGPT (Pham et al., 2024a) generate interpretable, natural-language

topics with LLMs, while CWTM (Fang et al., 2024) integrates contextual BERT embeddings into neural topic models for improved coherence. In multimodal settings, PromptMTopic (Prakash et al., 2023) combines textual and visual cues via LLMs to extract culturally aware topics from memes via prompting. More broadly, LVLMs offer unified representations for image–text pairs. However, their application to multimodal topic modeling remains underexplored. To address this, CEMTM leverages LVLMs, using the final token as a compact and aligned multimodal embedding, enabling efficient and interpretable topic discovery by using LLM's pretrained knowledge, without separate modality encoders or prompting.

3 Method

CEMTM is designed to perform soft topic modeling over long, multimodal documents. As shown in Figure 1, CEMTM processes both text and image inputs through an LVLM to produce contextualized token embeddings for both image and text tokens, learns importance-aware topic vectors, and reconstructs semantic document-level representations as supervision. We present our approach in three parts: document preprocessing (Section 3.1), model training (Section 3.2), and topic extraction (Section 3.3).

3.1 Preprocessing

Each document in the corpus contains both textual content and one or more associated images. Prior to training, we apply the following preprocessing steps. We begin with text cleaning, where we apply standard NLP preprocessing to remove punctuation, normalize casing, and eliminate irrelevant tokens (e.g., HTML tags). Following this, we perform vocabulary construction by tokenizing all documents and building a fixed vocabulary $\mathcal{V}$ that retains the most frequent words while discarding stop-words and rare terms. For the image processing step, all associated images are resized and formatted to ensure compatibility with the input requirements of the vision-language model.

3.2 Model Training

We use VLM2Vec (Jiang et al., 2025), a fine-tuned version of LLaVA-Next-7B (Liu et al., 2024), to encode each document's text and image content into contextualized representations. Our approach is motivated by the hypothesis that while document
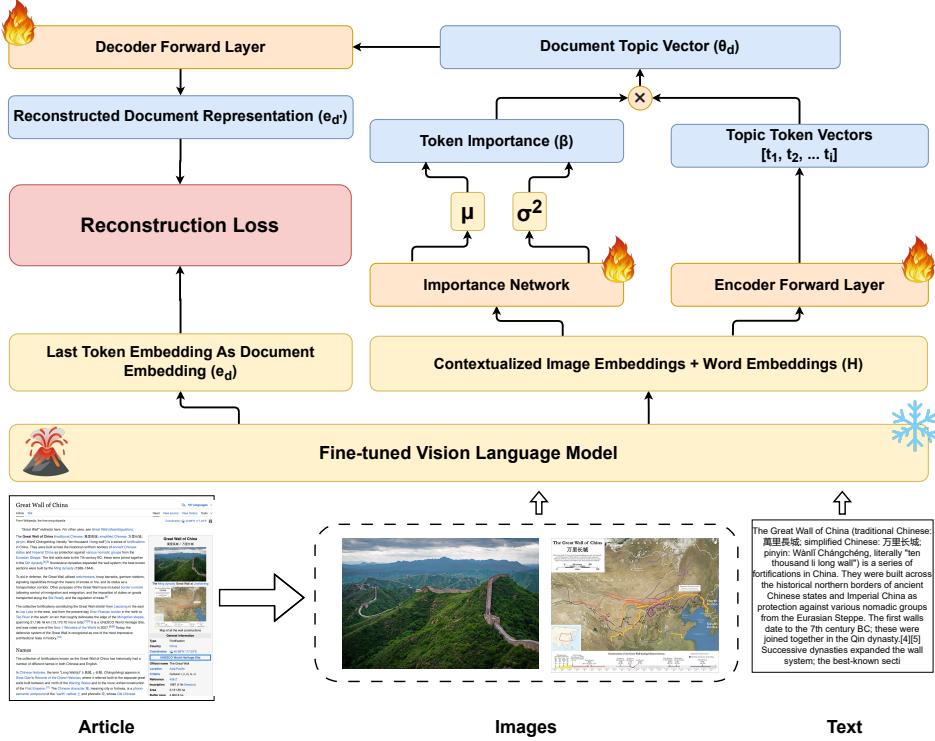
Figure 1: Overall architecture of CEMTM. Articles containing both text and images are encoded using a fine-tuned vision–language model to produce contextualized embeddings. During training, only the **decoder forward layer**, **encoder forward layer**, and **importance network** are fine-tuned, while the underlying vision–language backbone remains frozen. The model learns to construct document topic vectors by weighting token embeddings through the importance network, with reconstruction loss guiding optimization.

embeddings encode rich semantic information, using them alone to infer topic distributions prevents access to vocabulary-level topic-word associations, limiting interpretability.

We begin by considering the approach of inferring latent document-topic vectors from document embeddings. Let $\mathbf{e}_d \in \mathbb{R}^D$ be the embedding of a document d obtained from an LVLM, where D denotes the dimensionality of the embedding space. A straightforward method would use the document embedding vector to generate the topics. However, this formulation lacks a way to associate topics with specific words, since it bypasses vocabulary-level granularity. To address this, we instead extract contextualized token embeddings from the document:

$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N] \in \mathbb{R}^{N \times D}$$

where N is the number of textual tokens and visual patches in the document. Each $\mathbf{h}_i$ corresponds to a context-dependent representation of a token or an image patch. Each contextual embedding $\mathbf{h}_i$ is projected into the topic space using a **learnable encoder** with weight $\mathbf{W}_t \in \mathbb{R}^{D \times K}$, where K denotes

the number of latent topics, as follows:

$$\mathbf{t}_i = Softmax(\mathbf{h}_i \mathbf{W}_t) \in \mathbb{R}^K$$

We interpret $\mathbf{t}_i = p(z \mid \mathbf{h}_i)$ as the soft topic distribution for token i. However, not all tokens contribute equally to the semantic representation of a document. To model the relative importance of each token in shaping the document's semantics, we introduce a **learnable** importance network that predicts a stochastic weight for each token. The importance network consists of a transformer encoder followed by a feedforward projection layer. Given contextualized token embeddings $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$, the importance network outputs a mean and standard deviation for each token's importance score:

$$\mu_i, \sigma_i^2 = f_\theta(\text{Transformer}(\mathbf{H}))_i$$

$$\alpha_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

To produce normalized importance weights, we apply a softmax across the sampled values:

$$\boldsymbol{\beta} = Softmax([\alpha_1, \ldots, \alpha_N]) \in \mathbb{R}^N$$

Advantageously, the importance network also improves **interpretability**. The stochastic weights $\boldsymbol{\beta}$ provide an explicit estimate of how much each token or patch contributes to the document-level semantics. This makes it possible to identify salient tokens or regions and trace the evidence behind topic assignments, offering a more transparent view compared to standard attention mechanisms.

The document-topic vector is then computed by taking a weighted average of the token-level topic vectors:

$$\boldsymbol{\theta}_d = Softmax(\sum_{i=1}^{N} \beta_i \mathbf{t}_i)$$

The document-topic vector is then mapped into the embedding space using a **learnable decoder** with weight $\mathbf{W}_d \in \mathbb{R}^{K \times D}$, where K is the number of latent topics and D is the embedding dimension:

$$e_{d'} = \boldsymbol{\theta}_d \mathbf{W}_d \in \mathbb{R}^D$$

For supervision, we use the final token's hidden state from VLM2Vec as the reference document embedding (e_d). The model maps the predicted embedding ($e_{d'}$) to this target, optimized with a reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \text{MSE}(e_{d'}, e_d)$$

This objective helps ensure that the learned topics preserve the global semantics encoded by the vision-language model, resulting in more coherent and multimodally grounded topic representations.

To encourage sharp and interpretable importance scores, we add an entropy regularization term to the loss (Vulić and Mrkšić, 2018). This term penalizes high-entropy (i.e., overly uniform) distributions over the importance weights β_i, pushing the model to concentrate attention on a smaller subset of relevant elements. This promotes sparsity in the importance scores, making the model's decisions more focused and interpretable, which benefits both transparency and performance in reasoning tasks. The entropy regularization is defined as:

$$\mathcal{L}_{\text{ent}} = \sum_{i=1}^{N} \beta_i \log \beta_i$$

We also apply a KL divergence penalty between the predicted importance distribution $q(\alpha_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ and a standard normal prior $p(\alpha_i) =$

$\mathcal{N}(0, 1)$. This regularization keeps topic importance variables close to a standard Gaussian, reducing overfitting and promoting a smooth, balanced latent space (Jin et al., 2021). This is crucial in multimodal settings to avoid overconfident or modality-biased topic representations.

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{N} \left(\log \frac{1}{\sigma_i} + \frac{\sigma_i^2 + \mu_i^2 - 1}{2} \right)$$

The final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}$$

where λ_{ent} and λ_{KL} are hyperparameters that control the strength of entropy and KL regularization, respectively.

This formulation enables the model to learn a flexible, distribution-based importance mechanism over tokens, while ensuring that the topic vector faithfully reconstructs document-level semantics and supports interpretable word-topic associations.

3.3 Topic Extraction

Once the model is trained, we extract topic-word associations by aggregating token-level topic vectors for each word in the vocabulary. Let $w \in \mathcal{V}$ be a word and $\mathcal{I}_w$ the set of all positions where w appears in the corpus. We compute the aggregated topic vector for word w as:

$$\mathbf{t}_w = \frac{1}{Z_w} \sum_{i \in \mathcal{I}_w} \beta_i \mathbf{t}_i,$$

where $Z_w = \sum_{i \in \mathcal{I}_w} \beta_i$ ensures normalization. The topic score for word w in topic k is $\mathbf{t}_w^{(k)}$, which is guaranteed to be non-negative due to the softmax used in the importance distribution. To extract representative topic words, we rank all words $w \in \mathcal{V}$ by their value $\mathbf{t}_w^{(k)}$ for each topic k.

Topic words are extracted based on the vocabulary. For image patches, we associate each patch with its nearest word token in the embedding space and use that token in the aggregation step. This allows us to incorporate the semantic contribution of visual information while keeping the topic-word distributions interpretable.

4 Experiments and Results

We conduct extensive experiments to evaluate the effectiveness of our proposed model, CEMTM, on both topic modeling and its application to topic-guided few-shot retrieval for multimodal question

answering. We assess the quality of the extracted topics using standard coherence and diversity metrics, and demonstrate the utility of the learned document-topic vectors in improving few-shot example selection. Additionally, we analyze the sensitivity of the model to the underlying encoder and provide qualitative insights into the learned topics and retrieval behavior. Refer to Appendix B for hyperparameter and experimental settings.

4.1 Datasets

We evaluate CEMTM across a diverse set of multimodal and long-document datasets spanning encyclopedic, scientific, narrative, educational, and social domains. Table 1 summarizes the datasets used in this study. Among these, only WikiWeb2M and SPIQA provide explicit ground-truth topic labels, which we use for quantitative evaluation. For the remaining datasets, we assess topic quality using unsupervised metrics such as coherence and diversity.

Dataset	Domain	# Docs	Avg. Tokens	Avg. Images
WikiWeb2M	Encyclopedic	100,833	527	4.1
SPIQA	Scientific	697	1342	3.7
VIST	Narrative	50,000	152	5.0
TQA	Educational	410	1086	2.9
MSCOCO	Image Captions	30,000	13	1.0
T4SA	Social Media	30,000	15	1.0
FHM	Memes	10,000	9	1.0

Table 1: Summary of datasets used in our experiments.

4.2 Evaluation Metrics

We evaluate topic quality using five standard metrics: Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014), Word Embedding score (WE) (Fang et al., 2016), LLM score (Stammbach et al., 2023), Inverse Rank-Biased Overlap (I-RBO) (Terragni et al., 2021), and Topic Diversity (TD) (Dieng et al., 2020). NPMI and WE measure word-level coherence (co-occurrence and semantic similarity), the LLM score uses a language model to rate coherence (1–3 scale), and has strong correlation with human judgments (Stammbach et al., 2023). I-RBO and TD capture diversity, respectively via rank-aware dissimilarity and unique word coverage.

For datasets with gold topic labels (WikiWeb2M, SPIQA), we also report clustering-based metrics: Purity (Zhao and Karypis, 2001), Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002), which measure alignment between predicted
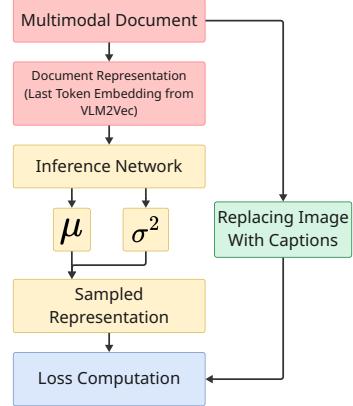
Figure 2: LVLM Zero-shot TM uses LVLM embeddings for better multimodal alignment and more meaningful topic vectors than Multimodal Zero-shot TM.

and true topic assignments.

4.3 Baselines

We compare CEMTM against a broad set of baselines spanning traditional, contextualized, and multimodal topic modeling. As a classical reference, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), trained with Gensim (Řehůřek and Sojka, 2010), models each document as a mixture of topics over a bag-of-words representation. More recent contextualized methods replace or augment bag of words (BoW) features with embeddings: ZeroshotTM (Bianchi et al., 2021a) uses SBERT embeddings (Reimers and Gurevych, 2019) for zero-shot topic modeling, and CombinedTM (Bianchi et al., 2021b) improves interpretability by concatenating SBERT with BoW. Similarly, CWTM (Fang et al., 2024) projects contextual token representations into a topic space and aggregates them using fixed or learned importance scores.

We also adapt TopicGPT (Pham et al., 2024a), which does not expose explicit topic-word distributions. To approximate them, we restrict the number of topics to K and use token-level soft assignments to construct interpretable topic-word vectors (see Appendix B for more details). Building on this, we introduce Multimodal TopicGPT, which incorporates both text and images at inference.

For other multimodal baselines, M3L-Contrast (Zosa and Pivovarova, 2022) leverages image–caption alignment to enforce consistent document-topic representations, while Multimodal ZeroshotTM (Gonzalez-Pizarro and Carenini, 2024a) extends ZeroshotTM by combining textual embeddings with vision encoder features. Finally,

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
LDA	.028	.095	2.40	.703	.953	.295	.131	.235	.022	.088	2.31	.717	.942	.299	.136	.244
CombinedTM	.039	.150	2.46	.696	.948	.317	.149	.258	.033	.140	2.39	.705	.940	.315	.148	.258
Zero-shot TM	.040	.172	2.51	.717	.966	.335	.149	.257	.036	.162	2.46	.731	.958	.331	.152	.263
CWTM	.052	.188	2.56	.714	.965	.347	.167	.275	.047	.177	2.51	.729	.957	.344	.168	.278
TopicGPT	.063	.212	2.59	.729	-	.378	.189	.288	.057	.201	2.55	.748	-	.377	.192	.294
M3L-Contrast	.065	.226	2.62	.744	.981	.386	.196	.298	.059	.215	2.59	.763	.973	.387	.199	.304
Multimodal Zero-shot TM	.071	.236	2.64	.756	-	.395	.204	.308	.062	.223	2.60	.776	-	.399	.206	.315
LVLM Zero-shot TM	.074	.246	2.65	.763	.990	.407	.213	.320	.065	.233	2.63	.785	.980	.411	.215	.326
Multimodal TopicGPT	.080	.255	2.67	.774	.993	.414	.224	.328	.071	.242	2.65	.798	.984	.419	.227	.335
CEMTM (ours)	**.088**	**.272**	**2.70**	**.792**	**.996**	**.435**	**.245**	**.351**	**.080**	**.258**	**2.68**	**.817**	**.987**	**.444**	**.251**	**.359**

Table 2: Comparison of topic modeling performance on WikiWeb2M (Burns et al., 2023) and SPIQA (Pramanick et al., 2024). We report coherence (NPMI, WE, LLM), diversity (TD), redundancy (I-RBO), and clustering metrics (Purity, ARI, NMI), averaged over $K = \{25, 50, 75, 100\}$ with three random seeds. CEMTM consistently outperforms all baselines. See Table 12 in Appendix D for detailed results for each K.

	VIST					TQA				
	NPMI	WE	LLM	TD	I-RBO	NPMI	WE	LLM	TD	I-RBO
LDA	.017	.077	2.23	.646	.935	.019	.081	2.25	.665	.940
CombinedTM	.024	.119	2.31	.637	.933	.028	.129	2.32	.652	.937
Zero-shot TM	.029	.138	2.38	.659	.949	.032	.151	2.39	.679	.955
CWTM	.036	.155	2.44	.656	.946	.041	.169	2.45	.675	.953
TopicGPT	.043	.179	2.47	.671	-	.050	.194	2.48	.692	-
M3L-Contrast	.044	.190	2.50	.681	.962	.052	.207	2.51	.705	.970
Multimodal Zero-shot TM	.048	.197	2.52	.687	.971	.056	.215	2.53	.716	.976
LVLM Zero-shot TM	.050	.208	2.54	.696	.974	.059	.226	2.55	.724	.977
Multimodal TopicGPT	.055	.216	2.56	.707	-	.064	.234	2.57	.736	-
CEMTM (ours)	**.062**	**.233**	**2.58**	**.723**	**.981**	**.071**	**.250**	**2.60**	**.752**	**.984**
	MSCOCO					T4SA				
	NPMI	WE	LLM	TD	I-RBO	NPMI	WE	LLM	TD	I-RBO
LDA	.016	.073	2.21	.618	.985	.012	.064	2.18	.597	.985
CombinedTM	.023	.117	2.28	.605	.984	.018	.105	2.25	.585	.978
Zero-shot TM	.027	.135	2.34	.629	.987	.023	.123	2.33	.610	.987
CWTM	.034	.153	2.40	.626	.987	.029	.142	2.38	.607	.988
TopicGPT	.042	.177	2.43	.642	-	.035	.164	2.42	.623	-
M3L-Contrast	.044	.189	2.46	.654	.990	.037	.175	2.45	.636	.991
Multimodal Zero-shot TM	.047	.198	2.48	.662	.992	.040	.182	2.46	.644	.992
LVLM Zero-shot TM	.050	.210	2.50	.670	.993	.043	.194	2.48	.652	.993
Multimodal TopicGPT	.055	.218	2.52	.682	-	.048	.202	2.50	.663	-
CEMTM (ours)	**.061**	**.233**	**2.54**	**.697**	**.995**	**.053**	**.218**	**2.52**	**.679**	**.995**

Table 3: Unsupervised topic quality on VIST (Huang et al., 2016), TQA (Kembhavi et al., 2017), MSCOCO (Lin et al., 2014), and T4SA (Vadicamo et al., 2017) using coherence (NPMI, WE, LLM), diversity (TD), and redundancy (I-RBO). Results are averaged over $K = \{25, 50, 75, 100\}$ with three random seeds. CEMTM outperforms all baselines. See Table 13 in Appendix D for detailed results for each K.

our proposed LVLM ZeroshotTM strengthens this approach by using embeddings from LVLMs, yielding more semantically grounded and better aligned multimodal topic vectors (as shown in Figure 2).

4.4 Quantitative Results

We evaluate the performance of CEMTM and baselines across a wide range of datasets, reporting both intrinsic topic quality metrics (e.g., NPMI, WE, LLM, TD, I-RBO) and extrinsic clustering metrics (Purity, ARI, NMI) when ground-truth labels are available. Results are averaged over four topic counts ($K = 25, 75, 100$), each run with

three random seeds.

Long-document and In Domain Performance. Table 2 presents results on WikiWeb2M and SPIQA, both of which consist of long, multimodal documents and include ground-truth topic annotations. CEMTM outperforms all baselines across every metric, demonstrating stronger topic coherence, higher diversity, and more accurate topic assignments. Notably, our model surpasses multimodal baselines like Multimodal TopicGPT and LVLM Zero-shot TM, while also being more efficient than methods like TopicGPT that require autoregressive

	FHM				
	NPMI	WE	LLM	TD	I-RBO
LDA	.005	.048	2.04	.530	.983
CombinedTM	.009	.087	2.10	.518	.975
Zero-shot TM	.014	.109	2.17	.543	.984
CWTM	.019	.128	2.22	.540	.986
TopicGPT	.025	.150	2.26	.554	.988
M3L-Contrast	.030	.169	2.34	.566	.990
Multimodal Zero-shot TM	.033	.177	2.36	.574	.992
LVLM Zero-shot TM	.039	.194	2.43	.590	.993
Multimodal TopicGPT	.043	.202	2.45	.601	.994
CEMTM (ours)	**.049**	**.217**	**2.47**	**.617**	**.995**

Table 4: Unsupervised topic quality on the FHM dataset (Kiela et al., 2020), which tests modeling under high image–text semantic gaps. We report coherence (NPMI, WE, LLM), diversity (TD), and redundancy (I-RBO), averaged over $K = \{25, 50, 75, 100\}$ with three seeds. CEMTM outperforms all baselines, highlighting the benefit of joint multimodal modeling. See Table 14 in Appendix D for detailed results for each K.

decoding or multiple forward passes (for topic generation and topic assignment). Unlike other models, CEMTM processes documents with multiple images in a single pass without repeated inference, offering both performance and scalability benefits.

Generalization Across Domains. Table 3 shows performance on four additional datasets—VIST, TQA, MSCOCO, and T4SA—that include both short and medium-length multimodal documents but lack ground-truth topic labels. Again, CEMTM achieves the best performance across all intrinsic metrics and datasets, highlighting its flexibility across domains including narratives (VIST), educational content (TQA), captioned images (MSCOCO), and social media posts (T4SA). These results indicate that the model generalizes well even beyond long-text scenarios.

Semantic Gap Analysis. A particularly challenging scenario in multimodal topic modeling arises when there is a semantic gap between images and text, as in the case of memes. Table 4 focuses on the Facebook Hateful Memes dataset, where there is a known semantic gap between images and their accompanying captions. This setting is particularly challenging for topic models that rely on textual content alone. The results show a clear separation between unimodal and multimodal models, with image-aware approaches consistently outperforming text-only counterparts. Furthermore, models that use large vision-language models (LVLMs), such as LVLM Zero-shot TM, Multimodal TopicGPT, and CEMTM, show the highest gains, sug-

Setting	SPIQA		TQA	
	METEOR	BERTScore-F1	Acc	F1-Macro
Zero-shot	26.3	67.48	84.87	83.79
3-shot Random Selection	27.4	68.92	85.36	84.28
3-shot Embedding Based Selection	28.7	70.11	86.09	85.12
3-shot Topic Based Selection	**31.3**	**72.76**	**87.31**	**87.03**

Table 5: Few-shot QA results on SPIQA and TQA test sets. Topic-based selection leads to the best performance across both datasets. For a detailed comparison of the performance of different topic models used for topic-based retrieval, refer to Table 11 in Appendix C.

gesting that better multimodal alignment significantly improves topic modeling in semantically ambiguous contexts. This validates the design of CEMTM, which leverages fine-tuned LVLM embeddings and a flexible importance-weighted fusion mechanism to capture cross-modal semantics effectively.

4.5 Improving Few-Shot Multimodal QA with Topic-Aware Retrieval

Beyond evaluating CEMTM on topic modeling tasks, we assess the utility of its learned document-topic vectors for improving few-shot multimodal question answering. Specifically, we use these topic vectors (with the number of topics set to $K = 50$) to retrieve in-context examples for prompting a QA model in a few-shot setting. We compare four retrieval strategies on the SPIQA and TQA test sets: (1) a zero-shot baseline, (2) random selection of 3 in-context examples, (3) embedding-based retrieval using cosine similarity over OpenAI's text-embedding-3-small[2], and (4) our topic-based retrieval using document-topic vectors produced by CEMTM. As shown in Table 5, topic-based selection significantly outperforms all other methods across all evaluation metrics, including METEOR and BERTScore on SPIQA, and accuracy and macro-F1 on TQA. This demonstrates that topic distributions learned by CEMTM capture high-level semantic structure that can guide effective example selection, providing relevant and diverse context without relying on direct surface similarity. These results highlight the potential of CEMTM beyond topic interpretability.

4.6 Qualitative Results

To further evaluate how CEMTM captures visually grounded semantics, we examine the Wikipedia article titled *Volcanic eruption*, which describes

---

[2] https://platform.openai.com/docs/models/text-embedding-3-small

Model	Topic Words Inferred from Text	Topic Words Inferred from Visual Patches
CWTM	eruption, magma, lava, ash, rock	—
Multimodal Zero-shot TM	eruption, lava, ash, magma, crater	plume, smoke, debris, slope, flow
LVLM Zero-shot TM	eruption, lava, ash, magma, vent	cloud, plume, lava flow, crater rim, tephra
CEMTM (ours)	eruption, lava, ash, magma, pyroclastic	plume, flow, cloud, fountain, tephra

Table 6: Predicted top topic words for the Wikipedia page *Volcanic eruption*, separated into text-derived vs. visual patch-derived words. The text columns intentionally show high overlap (e.g., *eruption*, *lava*, *ash*, *magma*) to reflect consistent lexical signals across models, with CEMTM additionally capturing specialized geological terminology such as *pyroclastic*. The visual columns highlight model-specific perceptual cues (e.g., *plume*, *flow*, *cloud*), showing how multimodal integration introduces eruption-specific semantics not present in text alone.

Query Page: *Saturn (planet)*	**Top Topic Words:** planet, ring, gas, orbit, atmosphere, moon, giant, solar, space, rotation
Random	*Barack Obama*, *Photosynthesis*, *Succulent plant*
Embedding-based	*Solar System*, *Mars*, *Astronomy*
Topic-based (CEMTM)	*Jupiter*, *Uranus*, *Gas giant*
Query Page: *French Revolution*	**Top Topic Words:** revolution, france, king, monarchy, liberty, citizens, republic, uprising, power, 1789
Random	*Harry Potter*, *Mount Everest*, *DNA replication*
Embedding-based	*American Revolution*, *Napoleon*, *History of France*
Topic-based (CEMTM)	*Reign of Terror*, *Louis XVI*, *Constitution of 1791*
Query Page: *Photosynthesis*	**Top Topic Words:** plant, sunlight, chlorophyll, carbon, dioxide, glucose, energy, leaf, oxygen, process
Random	*World War II*, *Twitter*, *Rome*
Embedding-based	*Cellular respiration*, *Chloroplast*, *Botany*
Topic-based (CEMTM)	*Light-dependent reactions*, *Carbon fixation*, *Thylakoid*

Table 7: Comparison of retrieval methods for Wikipedia pages. CEMTM yields more fine-grained, thematically aligned results by leveraging interpretable topic distributions.

types of volcanic eruptions, geological processes, and associated hazards. The page includes key images such as eruption plumes, lava flows, and ash clouds that visually differentiate between explosive and effusive eruptions, information that is often only implicitly mentioned or not described in detail in the text. Table 6 presents a comparison of top topic words predicted by CWTM (text-only), Multimodal Zero-shot TM, LVLM Zero-shot TM, and CEMTM. The text-only model generates general geological terms and omits eruption-specific visual cues. Multimodal Zero-shot TM incorporates visual features but lacks deep integration, leading to less coherent topic-word clusters. LVLM Zero-shot TM improves topic specificity, capturing visual elements like "plume" and "cloud," while CEMTM further refines this by predicting visually aligned and geologically grounded terms (e.g. "pyroclastic"). CEMTM benefits from fine-grained fusion of text and image semantics during training, and its reconstruction objective ensures visual information is preserved in the topic structure, something BoW-based models discard.

Table 7 qualitatively illustrates how CEMTM enhances semantic retrieval by leveraging interpretable document-topic vectors. For each query Wikipedia article, CEMTM retrieves thematically precise pages by comparing topic distributions, outperforming both random and embedding-based baselines. While embedding-based methods retrieve broadly related pages (e.g., *Mars* for *Saturn*), they often lack topical granularity. In contrast, CEMTM identifies highly specific, contextually aligned documents such as *Gas giant* or *Constitution of 1791*, grounded in the core semantic fields of the queries. This demonstrates that topic-based retrieval with CEMTM not only captures more interpretable signals but also better models thematic structure, making it particularly useful for few-shot prompting and corpus exploration.

5 Ablation Studies

5.1 Impact of Vision-Language Embedding Quality

To assess the effect of vision-language pretraining and fine-tuning, we compare several variants that adjust how VLM2Vec is used in CEMTM. As shown in Table 8, replacing the VLM2Vec version of LLaVA-Next-7B model, obtained by fine-tuning LLaVA-Next-7B, entirely with pretrained LLaVA-

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
CEMTM	**.088**	**.272**	**2.70**	**.792**	**.996**	**.435**	**.245**	**.351**	**.080**	**.258**	**2.68**	**.817**	**.987**	**.444**	**.251**	**.359**
Without Distribution As Importance Netowork	.087	.269	2.69	.789	.996	.432	.242	.348	.078	.255	2.68	.814	.987	.441	.248	.356
No VLM2Vec	.083	.260	2.67	.776	.994	.424	.231	.335	.074	.246	2.66	.797	.985	.429	.235	.342
VLM2Vec only for Word Embedding	.085	.265	2.68	.780	.994	.426	.234	.338	.075	.249	2.67	.801	.986	.432	.239	.346
VLM2Vec only for Document Embedding	.085	.266	2.68	.781	.995	.428	.235	.340	.076	.251	2.67	.802	.986	.434	.240	.348

Table 8: Ablation results on WikiWeb2M and SPIQA, showing the impact of using distribution-based importance modeling and fine-tuned VLM2Vec embeddings for word and document representations.

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
CEMTM	**.088**	**.272**	**2.70**	**.792**	**.996**	**.435**	**.245**	**.351**	**.080**	**.258**	**2.68**	**.817**	**.987**	**.444**	**.251**	**.359**
CEMTM w/o Image	.070	.228	2.64	.765	.981	.398	.205	.310	.062	.215	2.62	.785	.973	.400	.207	.318
CEMTM w/ Caption	.077	.246	2.66	.778	.988	.417	.221	.332	.068	.235	2.64	.800	.980	.425	.229	.340

Table 9: Modality ablation results on WikiWeb2M and SPIQA. **CEMTM w/o Image** indicates removing the image modality entirely, while **CEMTM w/ Caption** replaces the image modality with GPT-4o generated captions.

Next-7B results in the largest performance drop, particularly in document clustering metrics. This confirms that alignment-aware fine-tuned embeddings are crucial for accurate topic representation. Using VLM2Vec only for token embeddings or only for document embeddings results in intermediate performance: both help individually, but full use of VLM2Vec (as in the original model) provides the strongest gains. These results highlight the importance of semantically aligned, multimodal representations at both word and document levels. We further investigated the sensitivity of CEMTM across different LVLMs in Appendix A.

5.2 Role of Distributional Supervision in the Importance Network

We further evaluate the effect of modeling importance weights as samples from a learned Gaussian distribution, rather than as deterministic values. As shown in Table 8, removing this distributional supervision and replacing it with a simple softmax network leads to a consistent drop in performance across coherence (NPMI, WE, LLM), diversity (TD), and clustering metrics (Purity, ARI, NMI). This confirms that stochastic importance modeling not only improves robustness, but also helps the model better focus on semantically relevant tokens or image regions, ultimately yielding higher-quality and more interpretable topic structures.

5.3 Role of Visual Signals in Multimodal Topic Modeling

Table 9 presents the results of our modality ablation study on WikiWeb2M and SPIQA. We observe that removing the image modality (**w/o Image**) substantially degrades performance across all met-

rics, confirming the crucial role of visual signals in enhancing topic coherence and clustering quality. When substituting images with automatically generated captions (**w/ Caption**), the performance improves compared to removing images entirely, but it still falls short of the full model (**CEMTM**). This finding is consistent with prior work emphasizing that captions only provide partial information about visual content, whereas direct image features capture richer multimodal cues. Overall, these results highlight the importance of incorporating image representations directly, rather than relying solely on textual surrogates.

6 Conclusion

We presented **CEMTM**, an interpretable multimodal topic model designed to extract coherent topics from both short and long documents containing text and images. CEMTM leverages fine-tuned LVLM embeddings alongside a distributional attention mechanism, combining contextualized representations with a reconstruction-based training objective and importance-weighted fusion. This enables the model to capture document-level semantics while preserving interpretability. Evaluated on six benchmark datasets, CEMTM achieves a strong average LLM score of 2.61 out of 3 and a Purity score of 0.44, outperforming a broad range of unimodal and multimodal baselines. Ablation results further highlight the value of fine-tuned LVLMs and distributional supervision in guiding topic quality. Overall, CEMTM is a scalable, explainable framework that enables tasks like few-shot retrieval, multimodal summarization, and corpus-level topic analysis with efficiency and interpretability.

Limitations

While CEMTM demonstrates strong performance and scalability across diverse multimodal datasets, several limitations remain. First, the model relies heavily on pretrained LVLMs, which introduces significant computational overhead and requires access to large-scale GPU resources (See Appendix B for more information). This may limit the applicability of CEMTM in low-resource or real-time settings. Second, although the reconstruction objective aligns topic vectors with semantic document embeddings, this does not guarantee that each topic is fully disentangled or interpretable in isolation– particularly when documents cover overlapping concepts or when visual information is noisy or redundant. Additionally, our evaluation focuses on English-language datasets and does not explore multilingual or cross-cultural settings, where visual semantics and topic interpretability may differ significantly. Lastly, while the importance network encourages interpretability through attention sparsity, its learned weights are not explicitly validated against human judgments, leaving room for future work in explainability and user-in-the-loop topic refinement.

Ethical Considerations

Potential Risks   This research presents potential risks related to the use of real-world multimodal data, which may contain harmful biases or inaccuracies. To mitigate these risks, all experiments were conducted in controlled settings, and none of the resulting models were deployed in public-facing systems. Additionally, we carefully monitored model outputs during evaluation to ensure that no harmful content was propagated.

FHM Offensive Data   We used the Facebook Hateful Memes (FHM) dataset, which contains potentially offensive content, strictly for experimental purposes in this study. To minimize harm, we do not release any models trained on this dataset. This precaution ensures that any biased or harmful patterns present in the data are not disseminated or used beyond the limited scope of our research.

AI Assistance   AI tools were used during this project to assist with both writing and coding. Specifically, AI assistance supported drafting text, refining code structure, and improving clarity. However, all scientific contributions, including experimental design, analysis, and interpretation, were solely conducted by the authors to preserve research integrity.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

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, Online. Association for Computational Linguistics.

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. ACM.

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

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1057–1060.

Zheng Fang, Yulan He, and Rob Procter. 2024. CWTM: Leveraging contextualized word embeddings from BERT for neural topic modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4273–4286, Torino, Italia. ELRA and ICCL.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California. Association for Computational Linguistics.

Felipe Gonzalez-Pizarro and Giuseppe Carenini. 2024a. Neural multimodal topic modeling: A comprehensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12159–12172, Torino, Italia. ELRA and ICCL.

Felipe Gonzalez-Pizarro and Giuseppe Carenini. 2024b. Neural multimodal topic modeling: A comprehensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12159–12172.

Amin Hosseiny Marani and Eric P. S. Baumer. 2023. A review of stability in topic modeling: Metrics for assessing and techniques for improving stability. *ACM Comput. Surv.*, 56(5).

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and 1 others. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*.

Yuan Jin, He Zhao, Ming Liu, Lan Du, and Wray Buntine. 2021. Neural attention-aware hierarchical topic model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1052, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes.

In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10160–10171, Torino, Italia. ELRA and ICCL.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024a. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024b. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-Wei Lee. 2023.

Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 621–631.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *NeurIPS*.

Duangmanee Putthividhy, Hagai T. Attias, and Srikantan S. Nagarajan. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3408–3415.

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. 2021a. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, pages 33–45. Springer.

Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE*

International Conference on Computer Vision Workshops (ICCVW), pages 308–317.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis.

Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. 2014. Topic modeling of multimodal data: an autoregressive approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1370–1377.

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 Pivovarova. 2022. Multilingual and multimodal topic modelling with pretrained embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4037–4048, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Encoding Model Sensitivity

To evaluate the impact of the underlying vision-language encoder on CEMTM, we compare VLM2Vec variants fine-tuned from LLaVA-Next-7B (Li et al., 2024a) and QWen2VL-7B (Yang et al., 2024) (as described in Jiang et al. (2025)) with several non–fine-tuned baselines, including Phi-3.5-V (Abdin et al., 2024), LLaVA-1.6-7B (Li et al., 2024b), CLIP (Radford et al., 2021a), and BLIP2-OPT-7B (Li et al., 2023). As shown in Table 10, the choice of encoder has a clear effect on topic quality and clustering performance. QWen2VL-7B achieves the strongest results on both WikiWeb2M and SPIQA, while LLaVA-1.6-7B emerges as the most competitive non–fine-tuned baseline. These results highlight the importance of robust multimodal grounding for improving coherence and interpretability in CEMTM.

We also observe that VLM2Vec fine-tuning substantially improves document representations, enabling stronger topic coherence and clustering. Among the non–fine-tuned models, multimodal language models such as LLaVA-1.6 and Phi-3.5-V generally outperform vision-only encoders like CLIP, underscoring the advantage of joint vision–language reasoning in capturing corpus-level semantics.

B Experimental and Hyperparameter Settings

Experimental Settings All experiments were conducted using two **NVIDIA A100 80GB GPUs**. To account for variance in training, we report results averaged over 3 random seeds. This setup ensures consistency and robustness across different runs, especially when training large-scale models such as our proposed CEMTM and the multimodal baselines.

Hyperparameter Settings For our model, CEMTM, we use VLM2Vec as the encoder, based on a fine-tuned LLaVA-Next-7B. As detailed in Appendix A, we explore the impact of different LVLMs. All token embeddings are projected into a K-dimensional topic space. The importance network is a 2-layer Transformer (4 heads), followed by a feedforward layer predicting Gaussian token-level importance scores. The encoder forward layer is a 2-layer MLP with hidden size 512. We train with batch size 8, learning rate 2×10^{-5}, for 30 epochs using

Adam. Regularization weights are $\lambda_{\text{ent}} = 0.05$ and $\lambda_{\text{KL}} = 0.1$.

For baselines, we use public implementations when available. LDA is trained via Gensim with 100 passes and $\alpha = 0.01$. ZeroshotTM and CombinedTM use SBERT (all-MiniLM-L6-v2) with default settings from Bianchi et al. (2021a). TopicGPT and its multimodal variants are run with our modified version, limiting to K topics and assigning tokens sequentially to reflect topic preference. We then extract topic-word distributions by aggregating the token-topic assignments across the corpus, using soft alignment weights to represent each word's contribution to each topic. M3L-Contrast and Multimodal Zeroshot TM use CLIP ViT-B/32 (Radford et al., 2021b) for image features and SBERT (all-MiniLM-L6-v2) for text encoding. For text-only models (e.g., LDA, ZeroshotTM, CombinedTM, CWTM), we append GPT-4o-generated image captions to inputs to enable multimodal evaluation. All models use the same number of topics, tokenization, and document splits for fair comparison.

C Topic Models Comparison for Few-shot Retrieval

To better understand the role of topic models in guiding few-shot retrieval, we conduct a detailed comparison of different strategies for selecting in-context examples. The results in Table 11 reveal several important trends. First, simple baselines such as random selection or nearest-neighbor retrieval using document embeddings provide only modest improvements over the zero-shot setting. While these methods occasionally retrieve semantically similar examples, they lack the ability to capture higher-level topical coherence, which is crucial for complex multimodal QA tasks.

By contrast, topic-driven retrieval methods deliver more consistent and meaningful gains across datasets and metrics. Models such as CWTM and M3L-Contrast highlight the benefit of leveraging contextualized topic spaces, where representations capture recurring semantic patterns that extend beyond surface-level similarity. Extending this idea to multimodal topic models, Multimodal ZeroshotTM and our multimodal adaptation of TopicGPT further improve retrieval quality by incorporating both textual and visual signals. This demonstrates that aligning topics across modalities helps identify examples that are not only textually relevant but

	VLM2Vec	WikiWeb2M								SPIQA							
	Fine Tuned	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
LLava-Next-7B	✓	.087	.269	2.69	.789	.996	.432	.242	.348	.080	.258	2.68	.817	.987	.444	.251	.359
QWen2VL-7B	✓	**.093**	**.280**	**2.72**	**.796**	**.997**	**.444**	**.254**	**.361**	**.084**	**.269**	**2.71**	**.824**	**.991**	**.459**	**.263**	**.371**
Phi-3.5-V	✗	.083	.255	2.67	.777	.993	.414	.228	.332	.074	.244	2.65	.804	.984	.425	.235	.342
LLaVA-1.6-7B	✗	<u>.088</u>	<u>.272</u>	<u>2.70</u>	<u>.792</u>	<u>.996</u>	<u>.435</u>	<u>.245</u>	<u>.351</u>	<u>.080</u>	<u>.258</u>	<u>2.68</u>	<u>.817</u>	<u>.987</u>	<u>.444</u>	<u>.251</u>	<u>.359</u>
CLIP	✗	.081	.251	2.67	.774	.991	.410	.222	.328	.072	.241	2.65	.800	.982	.421	.231	.339
BLIP2-OPT-7B	✗	.083	.255	2.67	.777	.993	.414	.228	.332	.074	.244	2.65	.805	.984	.425	.235	.343

Table 10: Impact of the underlying LVLM encoder on CEMTM performance. We compare LoRA fine-tuned vision-language models—LLaVA-Next-7B, LLaVA-1.6-7B, QWen2VL-7B, Phi-3.5-V, CLIP, and BLIP—as backbone encoders for CEMTM. Results are reported on WikiWeb2M and SPIQA across topic coherence (NPMI, WE, LLM).

Setting	SPIQA		TQA	
	METEOR	BERTScore-F1	Accuracy	F1-Macro
Zero-shot	26.3	67.48	84.87	83.79
3-shot Random Selection	27.4	68.92	85.36	84.28
3-shot Embedding-Based Selection	28.7	70.11	86.09	85.12
3-shot Topic-Based (CWTM)	28.3	69.85	85.82	84.96
3-shot Topic-Based (M3L-Contrast)	28.9	70.22	86.09	85.18
3-shot Topic-Based (Multimodal Zero-shot TM)	29.4	70.63	86.23	85.43
3-shot Topic-Based (LVLM Zero-shot TM)	29.8	71.18	86.58	85.84
3-shot Topic-Based (Multimodal TopicGPT)	30.5	71.89	86.82	86.39
3-shot Topic-Based (CEMTM)	**31.3**	**72.76**	**87.31**	**87.03**

Table 11: Few-shot multimodal QA performance on SPIQA and TQA using various retrieval strategies for selecting 3 in-context examples. Topic-based retrieval with CEMTM consistently outperforms baselines across all metrics.

also visually coherent, which is particularly important in settings like SPIQA that require reasoning over mixed modalities. Among all topic-based approaches, CEMTM achieves the best performance. Its contextualized embeddings, projected into a coherent topic space, allow for more fine-grained retrieval that balances both semantic richness and cross-modal alignment. This enables the model to consistently select in-context examples that are well-suited to the target question, leading to measurable improvements in answer quality.

D Detailed Results For All Ks

We present the detailed results of various topic modeling approaches. Table 12 reports results across all K values (25, 50, 75, 100) for the WikiWeb2M and SPIQA datasets. Table 13 provides the corresponding results for the VIST, MSCOCO, T4SA, and TQA datasets. Table 14 shows the detailed performance of different topic modeling models on the FHM dataset across all K values.

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
K=25																
LDA	.031	.098	2.39	.708	.956	.306	.138	.240	.024	.090	2.30	.726	.944	.311	.142	.248
CombinedTM	.042	.154	2.45	.700	.950	.322	.155	.262	.035	.143	2.38	.715	.942	.324	.153	.263
Zero-shot TM	.042	.176	2.50	.724	.968	.340	.153	.261	.038	.166	2.45	.740	.961	.341	.157	.268
CWTM	.057	.193	2.55	.720	.967	.352	.172	.279	.049	.181	2.50	.739	.960	.354	.174	.282
TopicGPT	.068	.218	2.59	.736	.980	.384	.195	.292	.059	.205	2.54	.758	.972	.386	.198	.298
M3L-Contrast	.070	.232	2.62	.749	.982	.392	.202	.302	.061	.219	2.57	.772	.975	.396	.205	.308
Multimodal Zero-shot TM	.075	.240	2.63	.763	.990	.402	.210	.312	.064	.226	2.59	.785	.981	.408	.212	.318
LVLM Zero-shot TM	.078	.251	2.64	.770	.990	.414	.220	.324	.067	.237	2.62	.794	.982	.419	.221	.330
Multimodal TopicGPT	.085	.259	2.66	.782	.994	.421	.231	.331	.072	.245	2.64	.806	.985	.428	.233	.339
CEMTM	.092	.273	2.68	.796	.996	.438	.247	.351	.081	.262	2.67	.825	.988	.452	.256	.362
K=50																
LDA	.029	.096	2.40	.705	.954	.296	.132	.238	.023	.088	2.31	.720	.943	.302	.137	.245
CombinedTM	.040	.152	2.45	.698	.949	.319	.151	.261	.033	.141	2.39	.708	.941	.317	.149	.259
Zero-shot TM	.041	.174	2.51	.719	.967	.337	.150	.259	.037	.163	2.46	.733	.959	.334	.153	.265
CWTM	.053	.190	2.55	.717	.966	.349	.169	.277	.047	.179	2.51	.732	.958	.346	.169	.279
TopicGPT	.064	.214	2.59	.732	.979	.380	.192	.290	.058	.202	2.55	.751	.970	.378	.193	.295
M3L-Contrast	.066	.229	2.62	.746	.982	.389	.198	.300	.060	.217	2.58	.766	.974	.388	.200	.305
Multimodal Zero-shot TM	.072	.238	2.64	.758	.989	.398	.206	.310	.063	.224	2.60	.779	.980	.400	.207	.316
LVLM Zero-shot TM	.075	.248	2.65	.766	.990	.409	.216	.322	.066	.234	2.63	.788	.981	.412	.216	.327
Multimodal TopicGPT	.081	.257	2.67	.777	.993	.416	.227	.330	.072	.243	2.65	.800	.984	.420	.228	.336
CEMTM	.088	.271	2.69	.791	.996	.433	.244	.349	.081	.26	2.68	.82	.987	.445	.252	.360
K=75																
LDA	.027	.094	2.40	.702	.952	.291	.129	.233	.021	.087	2.32	.714	.941	.296	.134	.242
CombinedTM	.038	.149	2.46	.695	.947	.315	.147	.257	.032	.139	2.39	.701	.939	.312	.146	.256
Zero-shot TM	.039	.171	2.51	.714	.965	.334	.148	.255	.035	.161	2.46	.727	.957	.328	.151	.262
CWTM	.050	.186	2.56	.712	.964	.345	.166	.274	.046	.176	2.51	.726	.956	.341	.166	.276
TopicGPT	.061	.210	2.60	.727	.978	.376	.188	.287	.056	.199	2.56	.745	.969	.374	.19	.293
M3L-Contrast	.064	.224	2.63	.742	.981	.384	.195	.296	.058	.214	2.59	.760	.972	.384	.197	.303
Multimodal Zero-shot TM	.069	.234	2.64	.753	.989	.393	.202	.306	.062	.222	2.61	.773	.978	.397	.204	.314
LVLM Zero-shot TM	.072	.244	2.66	.761	.990	.404	.211	.318	.065	.232	2.63	.782	.980	.408	.213	.325
Multimodal TopicGPT	.078	.253	2.67	.772	.993	.412	.223	.326	.070	.241	2.66	.795	.983	.417	.225	.334
CEMTM	.085	.267	2.70	.787	.995	.430	.240	.347	.079	.257	2.69	.815	.987	.441	.250	.358
K=100																
LDA	.025	.092	2.41	.698	.951	.288	.126	.230	.020	.085	2.32	.709	.940	.289	.131	.239
CombinedTM	.037	.147	2.46	.692	.946	.311	.144	.253	.030	.137	2.40	.696	.938	.306	.143	.253
Zero-shot TM	.038	.169	2.52	.710	.964	.330	.144	.251	.034	.159	2.47	.722	.955	.323	.148	.258
CWTM	.049	.183	2.56	.708	.962	.341	.162	.270	.044	.173	2.52	.720	.954	.336	.163	.273
TopicGPT	.059	.207	2.60	.722	.977	.371	.183	.283	.054	.196	2.57	.739	.967	.369	.187	.289
M3L-Contrast	.062	.221	2.63	.737	.980	.379	.190	.293	.057	.211	2.60	.754	.971	.379	.194	.300
Multimodal Zero-shot TM	.067	.231	2.65	.748	.988	.388	.197	.303	.061	.219	2.61	.767	.977	.392	.201	.311
LVLM Zero-shot TM	.070	.242	2.66	.755	.989	.399	.206	.315	.063	.229	2.64	.776	.979	.404	.210	.322
Multimodal TopicGPT	.076	.251	2.68	.766	.992	.407	.217	.323	.069	.238	2.67	.789	.982	.413	.222	.331
CEMTM	.083	.266	2.70	.781	.995	.425	.238	.345	.078	.254	2.70	.808	.986	.438	.247	.356

Table 12: Comparison of topic modeling performance on WikiWeb2M and SPIQA. We report coherence (NPMI, WE, LLM), diversity (TD), redundancy (IRBO), and clustering metrics (Purity, ARI, NMI).

	VIST					TQA					MSCOCO					T4SA				
	NPMI	WE	LLM	TD	I-RBO	NPMI	WE	LLM	TD	I-RBO	NPMI	WE	LLM	TD	I-RBO	NPMI	WE	LLM	TD	I-RBO
K=25																				
LDA	.019	.080	2.23	.653	.937	.021	.083	2.24	.670	.942	.018	.075	2.21	.624	.986	.013	.066	2.18	.602	.985
CombinedTM	.026	.123	2.30	.644	.935	.030	.132	2.32	.658	.939	.025	.121	2.27	.611	.985	.020	.108	2.25	.590	.977
Zero-shot TM	.031	.142	2.38	.666	.951	.034	.155	2.39	.684	.957	.029	.138	2.34	.635	.987	.025	.127	2.32	.614	.987
CWTM	.039	.159	2.43	.664	.949	.043	.173	2.44	.681	.955	.037	.157	2.39	.633	.987	.031	.145	2.38	.612	.988
TopicGPT	.046	.183	2.47	.678	.961	.052	.199	2.48	.697	.968	.044	.180	2.43	.648	.990	.037	.167	2.41	.627	.990
M3L-Contrast	.047	.194	2.50	.688	.964	.054	.211	2.51	.711	.972	.046	.193	2.46	.661	.990	.039	.179	2.44	.640	.991
Multimodal Zero-shot TM	.051	.201	2.51	.694	.973	.058	.219	2.52	.722	.978	.050	.201	2.47	.669	.992	.042	.186	2.46	.648	.992
LVLM Zero-shot TM	.053	.212	2.53	.703	.975	.061	.230	2.55	.730	.979	.053	.213	2.50	.677	.993	.045	.197	2.48	.656	.993
Multimodal TopicGPT	.058	.220	2.55	.714	.978	.066	.238	2.57	.742	.982	.057	.221	2.51	.688	.994	.050	.205	2.50	.667	.994
CEMTM	.065	.237	2.58	.730	.981	.073	.255	2.59	.758	.986	.063	.237	2.54	.703	.995	.055	.221	2.52	.683	.995
K=50																				
LDA	.018	.078	2.23	.648	.936	.02	.081	2.24	.667	.940	.017	.074	2.21	.620	.986	.012	.065	2.18	.599	.985
CombinedTM	.025	.120	2.31	.639	.934	.028	.130	2.32	.654	.937	.024	.118	2.28	.607	.984	.019	.106	2.25	.587	.978
Zero-shot TM	.029	.139	2.38	.661	.950	.033	.152	2.39	.680	.955	.028	.136	2.34	.631	.987	.024	.124	2.32	.611	.987
CWTM	.037	.157	2.44	.659	.948	.041	.170	2.45	.677	.953	.035	.154	2.40	.628	.987	.029	.143	2.38	.609	.988
TopicGPT	.044	.180	2.47	.673	.959	.050	.195	2.48	.693	.967	.042	.178	2.43	.643	.990	.035	.165	2.42	.624	.990
M3L-Contrast	.045	.191	2.50	.683	.963	.052	.208	2.51	.707	.970	.044	.190	2.46	.656	.990	.037	.176	2.45	.637	.991
Multimodal Zero-shot TM	.049	.199	2.52	.690	.972	.057	.215	2.53	.718	.976	.048	.199	2.48	.664	.992	.040	.183	2.46	.645	.992
LVLM Zero-shot TM	.051	.209	2.54	.698	.974	.060	.227	2.55	.726	.978	.051	.211	2.50	.672	.993	.044	.195	2.48	.653	.993
Multimodal TopicGPT	.056	.217	2.56	.709	.977	.065	.234	2.57	.738	.981	.056	.219	2.52	.683	.994	.048	.203	2.50	.665	.994
CEMTM	.063	.234	2.58	.725	.981	.072	.251	2.60	.754	.985	.062	.234	2.54	.698	.995	.054	.219	2.52	.681	.995
K=75																				
LDA	.017	.076	2.23	.643	.934	.019	.080	2.25	.664	.939	.016	.072	2.21	.616	.985	.011	.063	2.18	.596	.985
CombinedTM	.023	.117	2.31	.635	.933	.027	.129	2.32	.651	.936	.022	.116	2.28	.603	.984	.018	.104	2.25	.584	.978
Zero-shot TM	.028	.136	2.38	.656	.948	.032	.151	2.39	.677	.954	.027	.133	2.34	.627	.986	.022	.122	2.33	.608	.987
CWTM	.035	.154	2.44	.654	.946	.041	.168	2.45	.674	.952	.034	.152	2.40	.624	.987	.028	.140	2.38	.606	.988
TopicGPT	.042	.177	2.48	.669	.958	.049	.193	2.48	.690	.965	.041	.175	2.44	.640	.989	.034	.163	2.42	.621	.990
M3L-Contrast	.043	.189	2.50	.679	.961	.052	.205	2.52	.703	.969	.043	.188	2.46	.652	.990	.036	.174	2.45	.634	.991
Multimodal Zero-shot TM	.047	.196	2.52	.685	.971	.055	.213	2.53	.714	.975	.046	.197	2.48	.660	.992	.039	.181	2.46	.642	.992
LVLM Zero-shot TM	.049	.207	2.54	.694	.973	.059	.225	2.55	.722	.977	.049	.209	2.50	.668	.993	.042	.193	2.49	.650	.993
Multimodal TopicGPT	.054	.215	2.56	.705	.976	.063	.233	2.57	.734	.980	.054	.217	2.52	.679	.994	.047	.201	2.50	.661	.993
CEMTM	.061	.232	2.58	.722	.980	.071	.249	2.60	.749	.984	.060	.232	2.54	.695	.995	.053	.217	2.53	.677	.995
K=100																				
LDA	.015	.074	2.23	.638	.933	.018	.078	2.25	.659	.938	.015	.071	2.21	.612	.985	.011	.062	2.18	.593	.985
CombinedTM	.022	.115	2.31	.630	.931	.026	.126	2.32	.646	.934	.021	.113	2.28	.599	.984	.016	.102	2.26	.581	.977
Zero-shot TM	.027	.134	2.38	.652	.946	.031	.148	2.39	.673	.953	.025	.131	2.35	.623	.986	.021	.119	2.33	.605	.987
CWTM	.033	.151	2.44	.649	.944	.039	.165	2.45	.670	.950	.032	.149	2.40	.620	.987	.027	.138	2.39	.602	.988
TopicGPT	.040	.174	2.48	.664	.956	.048	.190	2.49	.686	.964	.039	.173	2.44	.636	.989	.033	.161	2.42	.618	.99
M3L-Contrast	.042	.186	2.51	.674	.960	.050	.203	2.52	.699	.968	.041	.186	2.47	.649	.990	.035	.172	2.45	.631	.991
Multimodal Zero-shot TM	.046	.193	2.52	.681	.970	.054	.211	2.53	.710	.974	.045	.194	2.48	.657	.992	.038	.179	2.47	.639	.992
LVLM Zero-shot TM	.048	.204	2.55	.690	.973	.057	.222	2.56	.719	.975	.048	.206	2.51	.665	.992	.041	.191	2.49	.647	.992
Multimodal TopicGPT	.053	.213	2.56	.701	.975	.062	.230	2.58	.731	.979	.053	.215	2.52	.676	.993	.046	.199	2.51	.658	.993
CEMTM	.059	.229	2.59	.718	.980	.069	.247	2.60	.746	.983	.059	.230	2.55	.691	.994	.051	.215	2.53	.674	.995

Table 13: Comparison of topic modeling performance on VIST, TQA, MSCOCO, and T4SA. We report coherence (NPMI, WE, LLM), diversity (TD), and redundancy (I-RBO).

	FHM				
	NPMI	WE	LLM	TD	I-RBO
			K=25		
LDA	.006	.051	2.04	.535	.983
CombinedTM	.011	.091	2.10	.523	.975
Zero-shot TM	.016	.113	2.17	.548	.985
CWTM	.021	.132	2.21	.546	.986
TopicGPT	.027	.154	2.26	.559	.989
M3L-Contrast	.032	.173	2.33	.571	.990
Multimodal Zero-shot TM	.034	.181	2.35	.580	.992
LVLM Zero-shot TM	.041	.198	2.42	.595	.993
Multimodal TopicGPT	.045	.206	2.44	.607	.994
CEMTM (ours)	.051	.221	2.47	.622	.995
			K=50		
LDA	.005	.049	2.04	.531	.983
CombinedTM	.010	.089	2.10	.519	.975
Zero-shot TM	.015	.110	2.17	.545	.984
CWTM	.020	.129	2.22	.542	.986
TopicGPT	.026	.151	2.26	.556	.988
M3L-Contrast	.030	.170	2.33	.567	.990
Multimodal Zero-shot TM	.033	.179	2.35	.576	.992
LVLM Zero-shot TM	.039	.196	2.43	.592	.993
Multimodal TopicGPT	.044	.204	2.45	.603	.994
CEMTM (ours)	.050	.219	2.47	.618	.995
			K=75		
LDA	.004	.047	2.04	.528	.983
CombinedTM	.009	.086	2.10	.516	.975
Zero-shot TM	.013	.108	2.17	.541	.984
CWTM	.019	.126	2.22	.539	.985
TopicGPT	.024	.149	2.26	.552	.988
M3L-Contrast	.029	.167	2.34	.564	.990
Multimodal Zero-shot TM	.032	.176	2.36	.573	.991
LVLM Zero-shot TM	.038	.193	2.43	.588	.993
Multimodal TopicGPT	.043	.201	2.45	.599	.993
CEMTM (ours)	.048	.216	2.47	.615	.995
			K=100		
LDA	.004	.045	2.05	.525	.982
CombinedTM	.008	.083	2.11	.512	.974
Zero-shot TM	.012	.105	2.17	.538	.984
CWTM	.018	.123	2.22	.535	.985
TopicGPT	.023	.146	2.27	.549	.988
M3L-Contrast	.028	.165	2.34	.561	.990
Multimodal Zero-shot TM	.031	.173	2.36	.569	.991
LVLM Zero-shot TM	.037	.190	2.43	.584	.993
Multimodal TopicGPT	.041	.198	2.45	.595	.993
CEMTM (ours)	.047	.213	2.47	.611	.995

Table 14: Unsupervised topic quality on the FHM dataset, which tests modeling under high image–text semantic gaps. We report coherence (NPMI, WE, LLM), diversity (TD), and redundancy (I-RBO).