CEMTM: Contextual Embedding-based Multimodal Topic Modeling

Anonymous ACL submission

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. 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 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 true topic disentanglement, 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. 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 textual content and a set of associated images. This approach not only captures deeply aligned cross-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 token to the document-topic representation. CEMTM achieves strong empirical performance across six benchmark datasets, obtaining an average LLM coherence score of **2.61**, 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 outperforms strong baselines on topic quality and downstream tasks like few-shot 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. More broadly, LVLMs offer unified representations for image–text pairs. CEMTM builds on this by using the final token of an LVLM as a compact, aligned multimodal document embedding—enabling efficient and interpretable topic discovery by using LLM's pre-trained 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 vision tokens, learns importance-aware topic vectors, and reconstructs semantic document-level representations as supervision. We present our approach in three parts: document pre-processing, model training, and topic extraction.

3.1 Pre-processing

Each document in the corpus contains both textual content and one or more associated images. Prior to training, we apply the following pre-processing steps. We begin with text cleaning, where we apply standard NLP pre-processing 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 stopwords 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 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

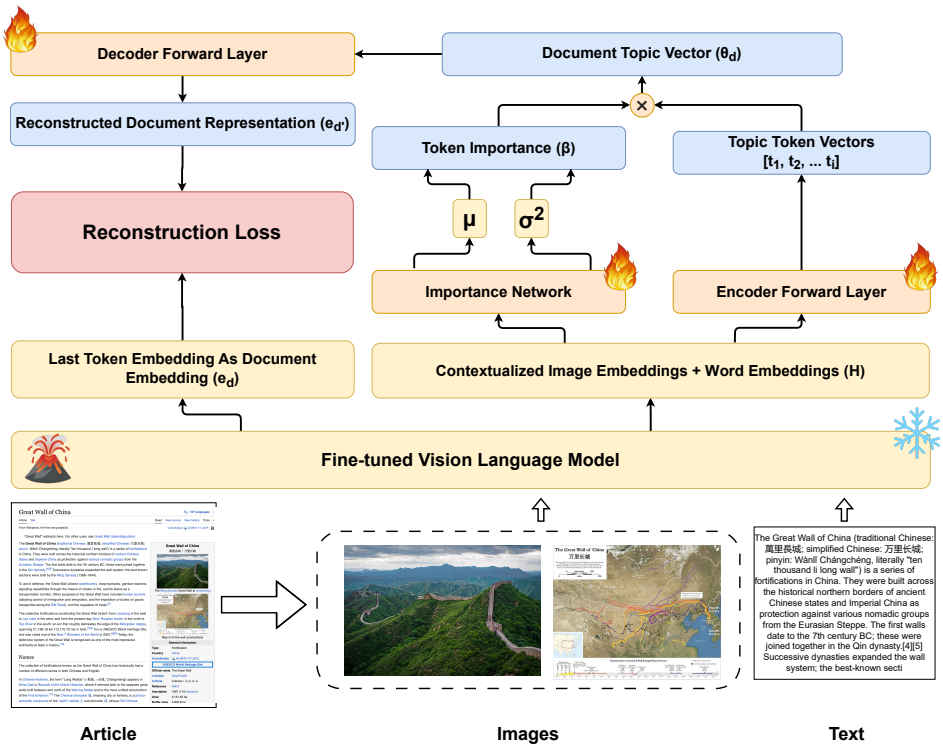


Figure 1: Architecture of CEMTM, our context-enhanced multimodal topic model. A document's text and images are encoded using VLM2Vec, a fine-tuned version of LLaVA-Next-7B, to produce contextualized token embeddings. These are processed by a forward encoder to produce topic-word vectors, and by an importance network (Transformer + feedforward layer) to compute a word importance distribution. The weighted topic vector is decoded into a document embedding and trained via a reconstruction loss against the reference embedding from the LVLM.

document d obtained from an LVLM. 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 $\mathbf{W}_t \in \mathbb{R}^{D \times K}$ as follows:

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

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:

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

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

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

To supervise training, we treat the final token's hidden state from VLM2Vec as the reference embedding for the entire document, as it encapsulates

high-level semantic information relevant to the document's content (e_d). The model learns to map the predicted document-topic vector to this reference embedding using a feedforward network. Training is guided by a reconstruction loss that minimizes the distance between the predicted embedding ($e_{d'}$) and the reference embedding, computed as:

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

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—beneficial for 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. This yields interpretable topic-word distributions while preserving soft assignments across the entire vocabulary.

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 and Appendix C for details on the evaluation metrics.

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.

4.1 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

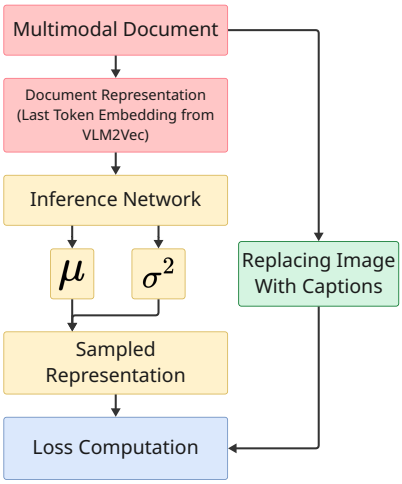


Figure 2: LVLM Zero-shot TM uses LVLM embeddings for better multimodal alignment and more meaningful topic vectors than Multimodal Zero-shot TM.

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.

4.2 Baselines

We compare CEMTM against a comprehensive set of baselines spanning traditional, contextualized, and multimodal topic modeling approaches.

LDA (Blei et al., 2003) is the classical Latent Dirichlet Allocation model, which we train using the implementation provided by Gensim (Řehřek and Sojka, 2010). It models each document as a mixture of latent topics over a bag-of-words (BoW) representation.

ZeroshotTM (Bianchi et al., 2021a) replaces BoW inputs with contextualized SBERT embeddings, enabling topic modeling in a zero-shot setting without explicit supervision.

CombinedTM (Bianchi et al., 2021b) extends ZeroshotTM by concatenating SBERT embeddings with BoW features to improve topic interpretability and alignment with text structure.

CWTM (Fang et al., 2024) combines contextualized word embeddings with a topic modeling framework. It projects contextual token representations into a topic space and aggregates them using fixed or learned importance scores to form a document-topic vector.

TopicGPT (Modified) (Pham et al., 2024a) could not be used directly, as it does not expose explicit topic-word distributions. To address this, we modify the original architecture by limiting the number of topics to K and collecting the soft topic assignments predicted for each token. These assignments are then used to build topic-word vectors by aggregating each word's contribution to different topics across the corpus. This allows us to approximate interpretable topic-word distributions, simulating a traditional topic model within the TopicGPT framework. See Appendix B for more details.

M3L-Contrast (Zosa and Pivovarova, 2022) is a contrastive multimodal topic model trained using image-caption alignment signals to enforce consistent document-topic representations across modalities.

Multimodal Zero-shot TM (Gonzalez-Pizarro and Carenini, 2024a) is a baseline that extends ZeroshotTM to the multimodal setting by incorporating mean-pooled image features from a vision encoder alongside text embeddings.

Multimodal TopicGPT is our extension of the modified TopicGPT, where both text and images are used during inference.

LVLM Zero-shot TM (Our introduced baseline) improves upon Multimodal Zero-shot TM by using embeddings from large vision-language models (LVLMs), resulting in better multimodal alignment and more semantically grounded topic vectors as shown in Figure 2.

4.3 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 ($K = 25, 50, 75, 100$), each run with three random seeds.

Long-document and Ground-truth Evaluation. 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 decoding or multiple forward passes (for topic gen-

	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 and SPIQA. 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 for detailed results for each K.

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

Table 3: Unsupervised topic quality on VIST, TQA, MSCOCO, and T4SA using coherence (NPMI, WE, LLM), diversity (TD), and redundancy (I-RBO). As these datasets lack ground-truth topics, only intrinsic metrics are shown. Results are averaged over $K = \{25, 50, 75, 100\}$ with three random seeds. CEMTM outperforms all baselines. See Table 13 for detailed results for each K.

eration 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. 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—suggesting 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.

4.4 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)

	FHM				
	NPMI	WE	LLM	TD	I-RBO
LDA	0.0051	0.0483	2.0482	0.5301	0.9832
CombinedTM	0.0098	0.0878	2.1074	0.5181	0.9752
Zero-shot TM	0.0145	0.1094	2.1756	0.5434	0.9847
CWTM	0.0199	0.1280	2.2222	0.5408	0.9860
TopicGPT	0.0255	0.1505	2.2674	0.5544	0.9887
M3L-Contrast	0.0302	0.1690	2.3410	0.5662	0.9905
Multimodal Zero-shot TM	0.0330	0.1776	2.3608	0.5749	0.9920
LVLM Zero-shot TM	0.0392	0.1947	2.4323	0.5902	0.9933
Multimodal TopicGPT	0.0438	0.2029	2.4529	0.6016	0.9940
CEMTM (ours)	**0.0496**	**0.2178**	**2.4758**	**0.6170**	**0.9953**

Table 4: 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), averaged over $K = \{25, 50, 75, 100\}$ with three seeds. CEMTM outperforms all baselines, highlighting the benefit of joint multimodal modeling. See Table 14 for detailed results for each K.

random selection of 3 in-context examples, (3) embedding-based retrieval using cosine similarity over OpenAI's `text-embedding-3-small`[1], 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.

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 E.

4.5 Qualitative Results

To further evaluate how CEMTM captures visually grounded semantics, we examine the Wikipedia article titled *Volcanic eruption*, which describes 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 "lava," 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. See Appendix D for more qualitative examples.

Model	**Top Predicted Topic Words**
CWTM	magma, rock, tectonic, energy, pressure
Multimodal Zero-shot TM	volcano, eruption, lava, mountain, damage
LVLM Zero-shot TM	lava, ash, crater, explosion, plume
CEMTM (ours)	eruption, plume, lava, pyroclastic, explosive

Table 6: Predicted top topic words for the Wikipedia page *Volcanic eruption*. While CWTM misses visual distinctions, LVLM-based models—especially CEMTM—capture eruption-specific visual-semantic features like "plume" and "pyroclastic."

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.

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

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.

	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.

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 VLM2Vec fine-tuned of LLaVA-Next-7B entirely with pre-trained LLaVA-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.

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 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 offers a scalable and explainable solution for downstream tasks such as few-shot retrieval, multimodal summarization, and corpus-level topic analysis.

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

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.

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.

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.

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.

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.

Duangmanee Putthividhy, Hagai T. Attias, and Srikantan S. Nagarajan. 2010. Topic regression multimodal 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, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

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.

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.

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 study the effect of the underlying vision-language encoder on CEMTM's performance, we compare different VLM2Vec variants, each LoRA fine-tuned from a distinct vision-language model: LLaVA-Next, QWen2VL-7B, and Phi-3.5-V (as described in Jiang et al. (2025)). As shown in Table 9, the choice of encoder significantly influences topic quality and clustering performance. QWen2VL-7B consistently yields the best results across both WikiWeb2M and SPIQA, highlighting its superior multimodal alignment and representation capabilities. These findings underscore the importance of strong vision-language grounding in improving topic coherence and downstream interpretability within CEMTM.

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., 2021) 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 Evaluation Metrics

We assess the intrinsic quality of the inferred topics using five widely adopted 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 measures co-occurrence of topic words within the corpus, while WE computes their pairwise similarity in a semantic embedding space. The LLM score uses a language model to assess topic coherence on a scale of 1–3, showing strong correlation with human judgments. I-RBO evaluates topic diversity by measuring rank-aware dissimilarity between all topic pairs, encouraging non-redundant topic representations. TD, on the other hand, quantifies the proportion of unique words across all topics, providing a lightweight estimate of overall diversity.

For datasets with ground-truth topic labels (WikiWeb2M and SPIQA), we can also apply clustering-based metrics to evaluate the alignment between predicted and true topic assignments. We report Purity (Zhao and Karypis, 2001), which evaluates the best matching between predicted and gold clusters based on harmonic precision and recall; Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), which accounts for all pairwise agreements corrected for chance; and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002), which quantifies the mutual dependence between the predicted and reference labels.

D More Qualitative Analysis

We analyze model predictions on a meme from the Facebook Hateful Memes (FHM) dataset, where the semantic meaning emerges from the interplay

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
LLava-Next	0.0875	0.2697	2.6978	0.7893	0.9960	0.4322	0.2427	0.3483	0.0800	0.2585	2.6893	0.8174	0.9876	0.4444	0.2517	0.3596
QWen2VL-7B	**0.0932**	**0.2800**	**2.7255**	**0.7964**	**0.9973**	**0.4446**	**0.2547**	**0.3612**	**0.0849**	**0.2694**	**2.7104**	**0.8244**	**0.9917**	**0.4594**	**0.2632**	**0.3717**
Phi-3.5-V	0.0834	0.2559	2.6762	0.7779	0.9931	0.4143	0.2284	0.3322	0.0748	0.2447	2.6579	0.8047	0.9843	0.4250	0.2353	0.3429

Table 9: Impact of the underlying LVLM encoder on CEMTM performance. We compare three LoRA fine-tuned vision-language models—LLaVA-Next, QWen2VL-7B, and Phi-3.5-V—as the backbone encoders for CEMTM. Results are reported on WikiWeb2M and SPIQA across topic coherence (NPMI, WE, LLM).

between text and image. In Table 10, we compare topic predictions from CWTM (text-only), Multimodal Zero-shot TM, LVLM Zero-shot TM, and CEMTM. CWTM focuses only on surface-level emotional or relationship cues found in the caption, while Multimodal Zero-shot TM adds vague visual context, but fails to resolve the irony. LVLM Zero-shot TM captures betrayal-related semantics more effectively. CEMTM produces the most grounded and expressive topic—capturing irony, betrayal, and emotional conflict—thanks to its reconstruction-guided alignment and importance-weighted multimodal fusion.

E Topic Models Comparison for Few-shot Retrieval

To further analyze the utility of topic-based document representations for in-context example selection, we provide a detailed comparison of few-shot retrieval performance across several baselines in Table 11. While random and embedding-based methods provide marginal improvements over the zero-shot baseline, topic-based retrieval methods offer consistent gains. CEMTM achieves the strongest performance among topic-driven models, demonstrating the effectiveness of its topic distributions in selecting relevant and semantically rich examples for prompting multimodal QA models.

F 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.

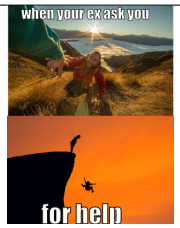
Meme (FHM Example)	Predicted Topic Words
	**CWTM:** breakup, sad, fall, help, push **Multimodal Zero-shot TM:** fall, person, ask, sky, support **LVLM Zero-shot TM:** betrayal, revenge, push, cliff, scream **CEMTM:** betrayal, help, irony, falling, manipulation

Table 10: Qualitative comparison of topic predictions on a meme from the FHM dataset. While CWTM is provided with an image caption, its lack of deep multimodal integration limits its ability to capture the intended meaning. Multimodal Zero-shot TM incorporates both modalities but encodes them separately, leading to less coherent semantics. LVLM Zero-shot TM improves visual grounding, and CEMTM further enhances interpretability by aligning cross-modal cues through joint reconstruction and attention-weighted fusion.

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.

	WikiWeb2M								SPIQA							
	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI	NPMI	WE	LLM	TD	I-RBO	Purity	ARI	NMI
	K=25															
LDA	0.0313	0.0987	2.3960	0.7085	0.9563	0.3068	0.1387	0.2406	0.0247	0.0905	2.3089	0.7265	0.9443	0.3114	0.1421	0.2489
CombinedTM	0.0429	0.1548	2.4519	0.7008	0.9503	0.3220	0.1556	0.2628	0.0351	0.1437	2.3842	0.7152	0.9428	0.3249	0.1538	0.2631
Zero-shot TM	0.0425	0.1760	2.5047	0.7246	0.9681	0.3403	0.1534	0.2619	0.0385	0.166	2.4556	0.7404	0.9612	0.3417	0.1575	0.2687
CWTM	0.0575	0.1936	2.5537	0.7204	0.9678	0.3524	0.1727	0.2796	0.0498	0.1819	2.5052	0.7397	0.9603	0.3542	0.174	0.2829
TopicGPT	0.0684	0.2185	2.5906	0.7361	0.9805	0.3847	0.1957	0.2929	0.0594	0.2053	2.5485	0.7581	0.9725	0.3866	0.1986	0.2988
M3L-Contrast	0.0701	0.2321	2.6208	0.7498	0.9827	0.3928	0.2025	0.3022	0.0612	0.219	2.5794	0.7729	0.9754	0.3961	0.2059	0.3087
Multimodal Zero-shot TM	0.0756	0.2403	2.6321	0.7630	0.9901	0.4024	0.2101	0.3125	0.0643	0.2264	2.5961	0.7855	0.981	0.4082	0.2124	0.3189
LVLM Zero-shot TM	0.0789	0.2511	2.6482	0.7708	0.9905	0.4140	0.2204	0.3243	0.0671	0.237	2.6228	0.7946	0.9821	0.4198	0.2217	0.3301
Multimodal TopicGPT	0.0853	0.2596	2.6651	0.7820	0.9941	0.4217	0.2312	0.3318	0.0728	0.2456	2.6479	0.8067	0.9853	0.4282	0.2336	0.3392
CEMTM	0.0923	0.2734	2.6892	0.7962	0.9963	0.4389	0.2473	0.3510	0.0816	0.2623	2.678	0.8256	0.9882	0.452	0.2564	0.3629
	K=50															
LDA	0.0295	0.0964	2.4045	0.7052	0.9540	0.2967	0.1324	0.2381	0.0231	0.0889	2.3157	0.7202	0.9431	0.3027	0.1372	0.2451
CombinedTM	0.0403	0.1523	2.4570	0.6981	0.9492	0.3194	0.1511	0.2612	0.0336	0.1418	2.3916	0.7087	0.9417	0.3178	0.1492	0.2598
Zero-shot TM	0.0412	0.1742	2.5103	0.7197	0.9675	0.3376	0.1508	0.2597	0.0372	0.1638	2.4625	0.7339	0.9596	0.334	0.1536	0.2652
CWTM	0.0532	0.1904	2.5599	0.7170	0.9660	0.3490	0.1693	0.2775	0.0479	0.1792	2.512	0.7325	0.9588	0.3461	0.1692	0.2791
TopicGPT	0.0641	0.2148	2.5967	0.7325	0.9796	0.3802	0.1923	0.2908	0.058	0.2025	2.5557	0.7514	0.9709	0.3785	0.1933	0.2956
M3L-Contrast	0.0667	0.2292	2.6275	0.7469	0.9821	0.3894	0.1989	0.3004	0.0602	0.2171	2.5868	0.7663	0.9742	0.3883	0.2007	0.3057
Multimodal Zero-shot TM	0.0721	0.2383	2.6402	0.7588	0.9897	0.3981	0.2066	0.3102	0.0635	0.2244	2.6037	0.779	0.9802	0.4007	0.2073	0.3162
LVLM Zero-shot TM	0.0754	0.2487	2.6563	0.7661	0.9902	0.4095	0.2163	0.3226	0.0664	0.2349	2.6304	0.7882	0.9815	0.4125	0.2165	0.3275
Multimodal TopicGPT	0.0816	0.2572	2.6720	0.7776	0.9937	0.4168	0.2276	0.3302	0.0721	0.2435	2.6554	0.8005	0.9847	0.4208	0.2284	0.3365
CEMTM	0.0887	0.2715	2.6961	0.7917	0.9960	0.4338	0.2441	0.3493	0.081	0.26	2.6856	0.82	0.9879	0.4456	0.2527	0.3608
	K=75															
LDA	0.0273	0.0942	2.4098	0.7020	0.9524	0.2913	0.1291	0.2334	0.0216	0.0871	2.3203	0.7146	0.9419	0.296	0.1341	0.2426
CombinedTM	0.0386	0.1492	2.4635	0.6952	0.9477	0.3155	0.1477	0.2573	0.0322	0.1397	2.3984	0.7017	0.9399	0.3122	0.1465	0.2566
Zero-shot TM	0.0398	0.1711	2.5160	0.7143	0.9659	0.3342	0.1482	0.2554	0.0359	0.1614	2.4699	0.7276	0.9577	0.3288	0.151	0.262
CWTM	0.0509	0.1863	2.5644	0.7124	0.9641	0.3457	0.1663	0.2749	0.0461	0.1764	2.5187	0.7262	0.9568	0.3413	0.1662	0.2764
TopicGPT	0.0619	0.2105	2.6012	0.7276	0.9783	0.3760	0.1884	0.2872	0.0565	0.1996	2.5621	0.7454	0.9693	0.3744	0.19	0.2931
M3L-Contrast	0.0642	0.2243	2.6325	0.7420	0.9814	0.3841	0.1950	0.2967	0.0588	0.2149	2.5933	0.7607	0.9727	0.3844	0.1976	0.3034
Multimodal Zero-shot TM	0.0694	0.2340	2.6454	0.7538	0.9892	0.3933	0.2021	0.3067	0.0622	0.2223	2.6102	0.7734	0.9788	0.3971	0.2043	0.314
LVLM Zero-shot TM	0.0728	0.2442	2.6615	0.7611	0.9900	0.4049	0.2117	0.3189	0.0651	0.2326	2.6369	0.7828	0.9803	0.4089	0.2136	0.3252
Multimodal TopicGPT	0.0781	0.2531	2.6768	0.7724	0.9931	0.4127	0.2230	0.3266	0.0708	0.2413	2.6619	0.7952	0.9837	0.4173	0.2255	0.3343
CEMTM	0.0854	0.2676	2.7008	0.7874	0.9959	0.4300	0.2408	0.3475	0.0794	0.2576	2.6918	0.815	0.9873	0.4418	0.2501	0.3582
	K=100															
LDA	0.0258	0.0924	2.4150	0.6987	0.9512	0.2881	0.1263	0.2300	0.0203	0.0855	2.3248	0.7092	0.9404	0.2892	0.1312	0.2398
CombinedTM	0.0375	0.1473	2.4690	0.6922	0.9460	0.3114	0.1440	0.2539	0.0309	0.1376	2.4051	0.6963	0.9382	0.3067	0.1439	0.2531
Zero-shot TM	0.0386	0.1695	2.5217	0.7108	0.9644	0.3301	0.1448	0.2511	0.0347	0.1591	2.4773	0.7223	0.9555	0.3231	0.1483	0.2587
CWTM	0.0491	0.1835	2.5691	0.7083	0.9628	0.3415	0.1623	0.2708	0.0446	0.1735	2.5271	0.7208	0.9542	0.336	0.1636	0.2734
TopicGPT	0.0599	0.2071	2.6060	0.7227	0.9774	0.3715	0.1832	0.2834	0.0549	0.1967	2.5702	0.7395	0.9678	0.3697	0.1871	0.2899
M3L-Contrast	0.0623	0.2214	2.6378	0.7372	0.9809	0.3795	0.1901	0.2931	0.0573	0.2116	2.6023	0.7543	0.9715	0.3797	0.1944	0.3002
Multimodal Zero-shot TM	0.0674	0.2316	2.6506	0.7486	0.9888	0.3888	0.1971	0.3031	0.061	0.2197	2.6193	0.7673	0.9775	0.3925	0.201	0.3111
LVLM Zero-shot TM	0.0708	0.2425	2.6669	0.7558	0.9898	0.3996	0.2064	0.3155	0.0639	0.2297	2.6462	0.7768	0.9791	0.4046	0.2103	0.3226
Multimodal TopicGPT	0.0762	0.2518	2.6812	0.7669	0.9928	0.4071	0.2178	0.3234	0.0696	0.2384	2.6711	0.7894	0.9826	0.413	0.2222	0.3317
CEMTM	0.0835	0.2664	2.7049	0.7818	0.9958	0.4259	0.2385	0.3454	0.0781	0.2542	2.7019	0.8089	0.9869	0.4382	0.2476	0.3564

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	0.0191	0.0803	2.2321	0.6531	0.9375	0.0212	0.0831	2.2482	0.6705	0.9421	0.0183	0.0758	2.2102	0.6241	0.9862	0.0132	0.0667	2.1812	0.6023	0.9851
CombinedTM	0.0265	0.1234	2.3085	0.6443	0.9358	0.0305	0.1324	2.3243	0.6581	0.939	0.0256	0.1213	2.2784	0.6113	0.9854	0.0205	0.1089	2.2525	0.5907	0.9779
Zero-shot TM	0.0311	0.142	2.381	0.6669	0.9511	0.0348	0.1556	2.3921	0.6845	0.9576	0.0299	0.1389	2.3433	0.6354	0.9873	0.0252	0.1271	2.3261	0.6149	0.9873
CWTM	0.039	0.1598	2.4378	0.6641	0.949	0.0437	0.1738	2.4479	0.6812	0.9553	0.0372	0.1572	2.3985	0.6331	0.9878	0.0314	0.1455	2.381	0.6126	0.9882
TopicGPT	0.0463	0.1831	2.4749	0.6787	0.961	0.0521	0.1992	2.4843	0.6975	0.9688	0.0445	0.1807	2.4349	0.6486	0.9902	0.0373	0.1674	2.4178	0.6278	0.9903
M3L-Contrast	0.0477	0.194	2.5025	0.6883	0.9645	0.0543	0.2115	2.5147	0.7111	0.9723	0.0464	0.1932	2.4642	0.6611	0.9908	0.039	0.179	2.4475	0.6404	0.9913
Multimodal Zero-shot TM	0.0512	0.2018	2.5174	0.6945	0.973	0.0582	0.2194	2.5284	0.722	0.9782	0.0501	0.2016	2.4786	0.6694	0.9926	0.0423	0.1862	2.4636	0.6487	0.9925
LVLM Zero-shot TM	0.0535	0.2126	2.5397	0.7031	0.9756	0.0611	0.2307	2.553	0.7304	0.9794	0.053	0.2133	2.5009	0.6772	0.9934	0.0455	0.1975	2.4841	0.6563	0.9932
Multimodal TopicGPT	0.0583	0.2207	2.5586	0.714	0.9781	0.066	0.2386	2.5712	0.7423	0.9821	0.0579	0.2218	2.5191	0.6885	0.9942	0.0502	0.2058	2.502	0.6675	0.9941
CEMTM	0.0651	0.2372	2.5812	0.7302	0.9819	0.0736	0.2552	2.5991	0.7583	0.986	0.0639	0.237	2.5425	0.7037	0.9952	0.0559	0.2218	2.524	0.6835	0.9955
	K=50																			
LDA	0.0182	0.0781	2.2344	0.6487	0.9362	0.02	0.0815	2.2498	0.6671	0.9407	0.0174	0.0743	2.2127	0.6205	0.986	0.0125	0.0651	2.1839	0.5991	0.9852
CombinedTM	0.0251	0.1203	2.3106	0.6395	0.9344	0.0287	0.1302	2.3268	0.6543	0.9378	0.024	0.1189	2.281	0.6071	0.9847	0.0191	0.1061	2.2552	0.5874	0.9784
Zero-shot TM	0.0297	0.1395	2.3838	0.6618	0.9502	0.0332	0.1526	2.395	0.6807	0.9557	0.0284	0.1363	2.3462	0.6311	0.9871	0.024	0.1246	2.3291	0.6117	0.9872
CWTM	0.0375	0.1574	2.4409	0.6591	0.948	0.0419	0.1704	2.4508	0.6773	0.9535	0.0355	0.1548	2.4014	0.6284	0.9877	0.0299	0.1432	2.3841	0.6092	0.9881
TopicGPT	0.0445	0.1803	2.4784	0.6739	0.9595	0.0506	0.1956	2.4873	0.6938	0.9671	0.0427	0.1783	2.4376	0.6439	0.99	0.0358	0.1653	2.4209	0.6247	0.9902
M3L-Contrast	0.0459	0.1915	2.5065	0.6837	0.9632	0.0529	0.208	2.5178	0.7072	0.9708	0.0447	0.1908	2.467	0.6564	0.9907	0.0376	0.1767	2.4507	0.6375	0.9912
Multimodal Zero-shot TM	0.0494	0.199	2.5212	0.6901	0.9721	0.057	0.2157	2.5316	0.7181	0.9769	0.0484	0.1994	2.4815	0.6646	0.9925	0.0408	0.1839	2.4669	0.6459	0.9924
LVLM Zero-shot TM	0.0518	0.2098	2.5439	0.6985	0.9748	0.0601	0.2271	2.5562	0.7266	0.9781	0.0514	0.2112	2.504	0.6725	0.9932	0.0441	0.1952	2.4875	0.6536	0.9931
Multimodal TopicGPT	0.0564	0.2179	2.5624	0.7092	0.9775	0.0651	0.2349	2.5744	0.7387	0.9811	0.0562	0.2197	2.5223	0.6837	0.9941	0.0488	0.2035	2.5053	0.665	0.994
CEMTM	0.0632	0.2347	2.5855	0.7252	0.9817	0.0727	0.2513	2.6025	0.7542	0.985	0.0623	0.2349	2.5452	0.6989	0.9951	0.0544	0.2193	2.5275	0.6812	0.9954
	K=75																			
LDA	0.017	0.0762	2.2367	0.6434	0.9348	0.0192	0.0806	2.2513	0.6642	0.9394	0.0162	0.0727	2.2149	0.6164	0.9858	0.0119	0.0639	2.1865	0.5964	0.9853
CombinedTM	0.0236	0.1176	2.3123	0.6351	0.9331	0.0278	0.1291	2.3279	0.6511	0.9361	0.0225	0.1164	2.2837	0.6031	0.9845	0.018	0.1041	2.2582	0.5842	0.978
Zero-shot TM	0.0281	0.1363	2.3857	0.6569	0.9483	0.0323	0.151	2.3971	0.6779	0.9544	0.0271	0.1337	2.3492	0.6271	0.9869	0.0229	0.1221	2.332	0.6086	0.9871
CWTM	0.0356	0.1546	2.4439	0.6542	0.9464	0.0412	0.1686	2.4534	0.6743	0.9522	0.034	0.1523	2.4044	0.6246	0.9875	0.0286	0.1409	2.3871	0.606	0.988
TopicGPT	0.0425	0.1778	2.4816	0.6691	0.9581	0.0498	0.1934	2.4896	0.6903	0.9657	0.0412	0.1759	2.4401	0.6403	0.9898	0.0346	0.1632	2.4239	0.6215	0.9901
M3L-Contrast	0.0438	0.1892	2.5097	0.679	0.9619	0.0521	0.2059	2.5202	0.7034	0.9696	0.0432	0.1884	2.4698	0.6527	0.9906	0.0364	0.1747	2.4536	0.6345	0.9911
Multimodal Zero-shot TM	0.0473	0.1967	2.5246	0.6852	0.9712	0.0559	0.2137	2.5341	0.7141	0.9757	0.0468	0.1972	2.4844	0.6607	0.9923	0.0395	0.1818	2.4698	0.6429	0.9923
LVLM Zero-shot TM	0.0498	0.2075	2.5475	0.6942	0.9739	0.0591	0.2252	2.5592	0.7227	0.977	0.0499	0.2091	2.5072	0.6687	0.9931	0.0429	0.1933	2.4903	0.6505	0.993
Multimodal TopicGPT	0.0545	0.2159	2.5662	0.7053	0.9767	0.0639	0.2333	2.578	0.7349	0.9802	0.0546	0.2176	2.5256	0.6798	0.994	0.0475	0.2017	2.5082	0.6618	0.9939
CEMTM	0.0613	0.2323	2.5891	0.7221	0.9809	0.0714	0.2497	2.6061	0.7499	0.9842	0.0607	0.2327	2.5481	0.6953	0.995	0.0532	0.2175	2.5303	0.6777	0.9954
	K=100																			
LDA	0.0159	0.0745	2.2381	0.6389	0.9332	0.018	0.0789	2.2529	0.6598	0.9381	0.0151	0.0711	2.2173	0.6128	0.9855	0.0111	0.0623	2.1881	0.5932	0.9852
CombinedTM	0.0223	0.1153	2.3147	0.6307	0.9315	0.0266	0.1269	2.3295	0.6469	0.9349	0.0212	0.1139	2.2863	0.5993	0.9842	0.0168	0.1021	2.2613	0.581	0.9777
Zero-shot TM	0.0271	0.1342	2.3882	0.652	0.9462	0.0312	0.1483	2.3993	0.6738	0.9531	0.0258	0.1312	2.3523	0.6233	0.9867	0.0217	0.1196	2.3348	0.6056	0.987
CWTM	0.0339	0.1519	2.4464	0.6494	0.9441	0.0399	0.1658	2.4559	0.6704	0.9508	0.0325	0.1499	2.4074	0.6209	0.9873	0.0272	0.1385	2.3902	0.6029	0.9879
TopicGPT	0.0409	0.1746	2.4846	0.6645	0.9568	0.0486	0.1909	2.492	0.6866	0.9644	0.0397	0.1734	2.443	0.6369	0.9896	0.0334	0.1611	2.4268	0.6184	0.99
M3L-Contrast	0.0423	0.1862	2.5128	0.6745	0.9607	0.0509	0.2034	2.5224	0.6997	0.9682	0.0419	0.186	2.4727	0.649	0.9904	0.0352	0.1725	2.4564	0.6315	0.991
Multimodal Zero-shot TM	0.0461	0.1939	2.5278	0.6811	0.9703	0.0548	0.2111	2.5362	0.7106	0.9746	0.0455	0.1949	2.4874	0.657	0.9921	0.0383	0.1796	2.4726	0.6398	0.9922
LVLM Zero-shot TM	0.0486	0.2048	2.5507	0.6903	0.973	0.0579	0.2226	2.5613	0.7193	0.9759	0.0486	0.2069	2.5104	0.665	0.9929	0.0417	0.1911	2.4932	0.6474	0.9929
Multimodal TopicGPT	0.0533	0.2132	2.5695	0.7015	0.9759	0.0628	0.2308	2.5801	0.7314	0.979	0.0533	0.2154	2.5288	0.676	0.9939	0.0463	0.1995	2.511	0.6586	0.9938
CEMTM	0.0598	0.2293	2.5926	0.7182	0.9802	0.0698	0.247	2.6082	0.7461	0.983	0.0594	0.2306	2.551	0.6917	0.9949	0.0519	0.2151	2.5331	0.6744	0.9953

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	0.0061	0.051	2.0442	0.5352	0.9836
CombinedTM	0.0114	0.0917	2.1021	0.5236	0.9758
Zero-shot TM	0.016	0.1132	2.1716	0.5489	0.985
CWTM	0.0216	0.132	2.2182	0.5461	0.9863
TopicGPT	0.0273	0.1545	2.2634	0.5596	0.989
M3L-Contrast	0.032	0.173	2.3371	0.5715	0.9907
Multimodal Zero-shot TM	0.0349	0.1816	2.3568	0.5804	0.9921
LVLM Zero-shot TM	0.0412	0.1987	2.4283	0.5959	0.9934
Multimodal TopicGPT	0.0459	0.2069	2.449	0.6073	0.9941
CEMTM (ours)	0.0517	0.2218	2.4719	0.6229	0.9954
	K=50				
LDA	0.0054	0.0492	2.0469	0.5317	0.9833
CombinedTM	0.0102	0.0891	2.1056	0.5199	0.9754
Zero-shot TM	0.015	0.1106	2.1743	0.5452	0.9848
CWTM	0.0204	0.1293	2.2209	0.5425	0.9861
TopicGPT	0.026	0.1518	2.2661	0.5561	0.9888
M3L-Contrast	0.0307	0.1703	2.3397	0.5679	0.9906
Multimodal Zero-shot TM	0.0335	0.179	2.3595	0.5767	0.992
LVLM Zero-shot TM	0.0398	0.196	2.431	0.5921	0.9933
Multimodal TopicGPT	0.0445	0.2043	2.4516	0.6035	0.994
CEMTM (ours)	0.0502	0.2192	2.4745	0.6189	0.9953
	K=75				
LDA	0.0048	0.0473	2.0496	0.5284	0.9831
CombinedTM	0.0092	0.0864	2.1092	0.5162	0.975
Zero-shot TM	0.0139	0.1081	2.1769	0.5416	0.9846
CWTM	0.0193	0.1266	2.2235	0.539	0.9859
TopicGPT	0.0249	0.1491	2.2687	0.5526	0.9886
M3L-Contrast	0.0296	0.1676	2.3423	0.5644	0.9904
Multimodal Zero-shot TM	0.0323	0.1763	2.3621	0.573	0.9919
LVLM Zero-shot TM	0.0385	0.1933	2.4336	0.5883	0.9932
Multimodal TopicGPT	0.0431	0.2016	2.4542	0.5997	0.9939
CEMTM (ours)	0.0489	0.2165	2.477	0.615	0.9952
	K=100				
LDA	0.0042	0.0455	2.0522	0.5251	0.9829
CombinedTM	0.0083	0.0838	2.1127	0.5125	0.9747
Zero-shot TM	0.0129	0.1056	2.1795	0.538	0.9845
CWTM	0.0182	0.1239	2.2262	0.5354	0.9858
TopicGPT	0.0237	0.1464	2.2714	0.5491	0.9885
M3L-Contrast	0.0285	0.165	2.3449	0.561	0.9903
Multimodal Zero-shot TM	0.0312	0.1736	2.3647	0.5693	0.9918
LVLM Zero-shot TM	0.0371	0.1906	2.4362	0.5845	0.9931
Multimodal TopicGPT	0.0418	0.1989	2.4568	0.5959	0.9938
CEMTM (ours)	0.0475	0.2138	2.4796	0.6111	0.9951

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).