M3-HNTM: Hyperspherical Multimodal Topic Modeling with Symbolic and Contextual Evidence

Zhiwen Luo[1] **Dayu Guo**[1] **Manar Amayri**[2] **Nizar Bouguila**[2]
Zhixiang Li[1] **Wenchuan Zhang**[1] **Wentao Fan**[1*]
[1]Beijing Normal-Hong Kong Baptist University [2]Concordia University
zhiwenluo@bnbu.edu.cn w630201603@mail.bnbu.edu.cn
{manar.amayri, nizar.bouguila}@concordia.ca
{u430201707, t330201704}@mail.bnbu.edu.cn
wentaofan@bnbu.edu.cn

Abstract

Multimodal evidence can improve topic discovery, but dense auxiliary signals can make topics harder to interpret when they replace readable descriptors. This creates an evidence-role problem: symbolic signals should define what topics mean to a reader, while dense aligned signals should help infer which topics a document expresses. We introduce M3-HNTM, a hyperspherical multimodal neural topic model that separates these roles through a shared document-level latent direction. M3-HNTM utilizes a shared von Mises-Fisher (vMF) posterior for document-level topic inference and vMF-mixture symbolic decoders that represent topics as concentration-aware semantic regions. Text Bag-of-Words and speech Bag-of-Acoustic-Words provide reconstructable symbolic evidence, while aligned image-text embeddings enter posterior inference as contextual evidence and help preserve semantic structure. The model applies to image-text, speech-text, and image-speech-text corpora while keeping topics readable through symbolic descriptors. Experiments on five datasets show that M3-HNTM improves the coherence-diversity quality score over strong text-only, hyperspherical, and multimodal baselines. On SpokenCOCO-Tri, the full tri-modal model outperforms both image-text and speech-text variants, indicating that visual and acoustic signals provide complementary evidence for interpretable topic discovery.

1 Introduction

Topic modeling is most useful when it makes a collection readable: an analyst can inspect topic descriptors, see how documents combine recurring themes, and compare those themes across a corpus [1–3]. This interpretable interface separates topic models from generic document embeddings [4, 5]. Classical LDA provides this structure by representing each document as a mixture of topics, with each topic defined as a distribution over words [1]. Neural topic models keep the same interface while improving variational inference [6–8], embedding-based topic-word modeling [9], and contextualized document evidence [10, 11]; recent surveys review this line of work [4, 12]. The main difficulty is that the most useful evidence for topic inference may now come from pretrained models or non-text modalities, while the topic interface still needs to stay symbolic and easy to inspect.

Multimodal corpora expose this difficulty. Captions and posts can be sparse, images add visual context, and speech can contain acoustic regularities missed by transcripts alone [13–15]. Image-text topic models have improved contextual modeling through benchmarked baselines and evaluation criteria [14, 16]. A recent study [17] further shows that large vision-language model (LVLM) encoders

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

can provide richer context for topic discovery. Speech presents a parallel problem from a different angle. Topic pipelines that depend on Automatic Speech Recognition (ASR) can be unreliable for spontaneous, noisy, or low-resource speech [15]. At the same time, acoustic-word representations and learned speech embeddings capture recurring acoustic patterns, but these signals are rarely used as direct topic evidence in modern neural topic models [18, 19]. These advances point to the same root problem. As Figure 1 shows, symbolic count evidence such as Bag-of-Words (BoW) gives topics readable descriptor anchors, whereas dense image-text features help disambiguate context. Directly concatenating these signals weakens this distinction: the problem is not simply how to fuse modalities, but how to assign each evidence source the right role.

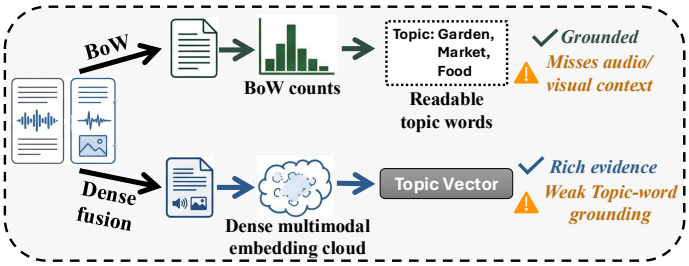


Figure 1: Motivating trade-off in multimodal topic evidence. The upper BoW path yields readable topic words but misses visual and acoustic context; the lower dense-fusion path captures richer context but weakens topic-word grounding.

When topic inference uses normalized pretrained embeddings such as CLIP [20], the need to keep evidence sources separate also has a geometric side. These representations are compared mainly by direction, so a latent document representation that combines symbolic counts with dense context should preserve angular semantics rather than arbitrary Euclidean length. A standard Gaussian posterior is a poor fit for this setting. Radial variation has no clear semantic meaning, but the posterior and its Kullback-Leibler (KL) penalty still need to model it [21]. When symbolic evidence is sparse, the KL penalty can pull the posterior toward the prior and weaken document-level topic signals [22–24]. A hyperspherical von Mises-Fisher (vMF) posterior reduces this mismatch by encoding semantic information in direction and using concentration to represent posterior certainty [2, 25, 26]. The decoder should follow the same geometry. Point-valued topic embeddings identify a topic direction but do not specify how broadly the topic should cover nearby words or acoustic units [8, 9]. In contrast, a geometry-consistent decoder can represent each topic as a concentration-aware semantic region [27–30].

Our key insight is that interpretable multimodal topic modeling should keep separate the evidence that makes a topic readable and the evidence that helps infer a document's topic mixture, while linking both through a shared directional latent geometry. We introduce M3-HNTM, a hyperspherical multimodal neural topic model for image-text, speech-text, and image-speech-text corpora. The model represents heterogeneous observations through a shared document-level latent direction on the unit hypersphere. BoW and Bag-of-Acoustic-Words (BoAW) provide reconstructable symbolic evidence for readable topic descriptors; BoAW converts continuous speech into countable acoustic-word evidence. Dense aligned image-text features are used instead to strengthen posterior inference and retain multimodal semantics without replacing the symbolic descriptor layer. A shared vMF posterior and vMF-mixture symbolic decoders place inference and reconstruction in the same directional geometry.

Contributions. M3-HNTM makes three contributions. First, it introduces a role-separated evidence design that keeps reconstructable BoW and BoAW descriptors separate from contextual dense features. Second, it provides a common hyperspherical formulation for image-text, speech-text, and image-speech-text topic modeling through one shared document latent direction. Third, it uses concentration-aware vMF-mixture decoders to represent topics as semantic regions rather than point embeddings. Experiments on five datasets show that M3-HNTM achieves the highest topic Quality among the compared baselines, and results on SpokenCOCO-Tri show that speech and visual evidence provide complementary signals.

2 Related Work

Neural topic models. Neural topic models retain the basic contract of classical topic modeling: documents are mixtures of topics, and topics can be inspected through their descriptors. They replace hand-designed inference with amortized variational inference and learned representations [12]. VAE-

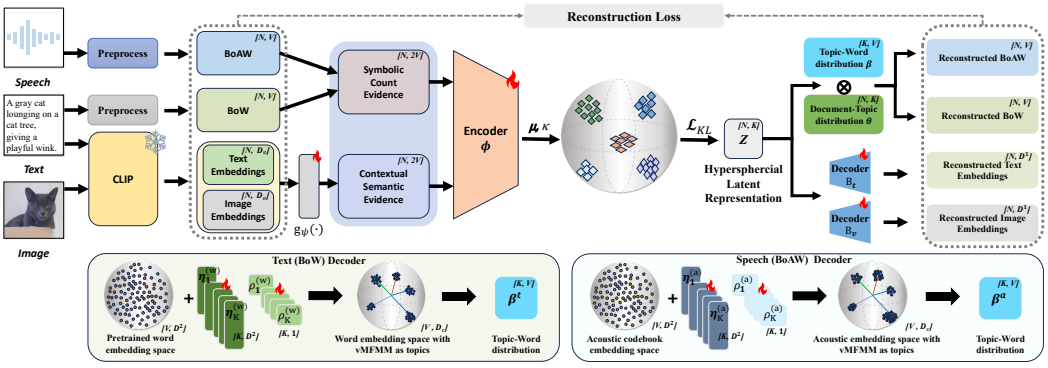


Figure 2: Architecture of M3-HNTM. Symbolic counts and calibrated dense context are encoded into a shared document latent direction $\mathbf{z}$, which is mapped to the document-topic mixture $\boldsymbol{\theta}$ by a temperature-controlled softmax. The same $\boldsymbol{\theta}$ drives vMF-mixture symbolic decoders for BoW and BoAW reconstruction; each topic's direction and concentration shape its distribution over words or acoustic codewords. Auxiliary dense heads reconstruct the raw aligned image-text embeddings.

based models, including NeuralLDA [7] and GSM [8], embedding-aware decoders such as ETM [9] and NVGMTM [31], contextualized models such as ZeroShotTM [11] and CombinedTM [10], and clustering-based methods such as BERTopic [32] have improved different parts of this pipeline. A related line of work questions whether this pipeline should be modeled in Euclidean space. Since L_2-normalized semantic embeddings are compared mainly by direction, hyperspherical VAEs [33], vMF-VAEs [34], spherical topic models such as SAM [35], SCTM [27], vONT [36], and S2WTM [28], as well as spherical mixture topic models [29], use directional priors, mixture distributions, or spherical transport to improve topic separation and coherence. These studies motivate our shared vMF posterior and vMF-mixture topic decoder. However, most of this work remains text-centered. It improves inference or topic-word geometry, but does not address how heterogeneous evidence from images, speech, and text should be assigned different roles within a single readable topic interface.

Multimodal topic models. Neural multimodal topic models target another limitation: sparse text often lacks semantic context, and non-text modalities can fill this gap. In image-text corpora, Multimodal-ZeroShotTM [14] adds visual features to ZeroShotTM [11], and Multimodal-Contrast uses cross-modal alignment [14]. Pretrained encoders like CLIP [20] make aligned representations practical. CEMTM [17] offers a strong recent LVLM-centered approach, using contextual multimodal token and patch encodings with importance weighting to improve image-text topic discovery. These methods show multimodal context is useful, but usually use non-text signals as fused features, alignment targets, or joint reconstruction targets. As a result, dense multimodal representations can carry much of the topic meaning, while the separation between auxiliary context and inspectable descriptors becomes less clear. Speech makes this issue clearer. Many speech-topic pipelines rely on ASR transcripts, but prior studies have criticized this dependency and explored audio-feature based alternatives [15]. Self-supervised speech features [19] and BoAW [18] retain recurring acoustic patterns in count-like form, but their use as reconstructable topic evidence in modern neural topic models remains limited. This motivates a stricter evidence-role design: M3-HNTM uses dense image-text features as contextual evidence, while keeping BoW and BoAW reconstruction as the symbolic topic interface, tying all evidence through a shared document latent direction.

3 Method

M3-HNTM assigns evidence roles before linking all observed modalities through the shared document latent direction in Figure 2. Symbolic counts define the readable topic interface, dense aligned features provide contextual evidence, and each available modality constrains the same document-topic mixture.

3.1 Preliminaries

vMF distribution. For a unit vector $\mathbf{y} \in \mathbb{S}^{d-1}$, the vMF distribution [33, 34] is

$$\mathrm{vMF}(\mathbf{y}; \boldsymbol{\epsilon}, \zeta) = C_d(\zeta) \exp(\zeta \boldsymbol{\epsilon}^\top \mathbf{y}), \qquad \|\boldsymbol{\epsilon}\|_2 = 1,\ \zeta \geq 0. \tag{1}$$

The mean direction $\boldsymbol{\epsilon}$ gives the semantic center, while the concentration ζ controls how tightly the distribution is concentrated around that center. M3-HNTM uses this distribution in two ways: the encoder infers a document-level hyperspherical latent direction, and the symbolic decoder represents each topic as a semantic region. For the document posterior, $d = K$, and we abbreviate $\mathrm{vMF}(\mathbf{z}_i; \boldsymbol{\mu}_i, \kappa_i)$ as $\mathrm{vMF}(\boldsymbol{\mu}_i, \kappa_i)$ when the random variable is clear.

BoAW construction. BoAW [18] converts a speech signal into a count vector over an acoustic codebook with the same size V as the text vocabulary. This shared size is a design choice for count reconstruction, not a one-to-one alignment between words and acoustic codewords. Given normalized frame-level speech embeddings [19] $\{\bar{\mathbf{s}}_{it}\}_{t=1}^{T_i}$ and normalized acoustic codewords $\mathbf{E}^{(a)} = \{\mathbf{e}_j^{(a)}\}_{j=1}^{V}$, we quantize each frame by angular similarity and then count the resulting assignments:

$$x_{ij}^{(a)} = \sum_{t=1}^{T_i} \mathbf{1}\left[j = \arg\max_{\ell \in \{1,\ldots,V\}} \bar{\mathbf{s}}_{it}^\top \mathbf{e}_\ell^{(a)}\right]. \tag{2}$$

This produces $\mathbf{x}_i^{(a)} \in \mathbb{R}_{\geq 0}^V$, allowing recurring acoustic patterns [15] to enter the same symbolic reconstruction path as BoW.

3.2 Evidence Representation

Depending on the dataset, document i contains text d_i and paired speech a_i, paired image v_i, or both. Let K denote the number of topics, V the shared count dimension for BoW and BoAW, D_c the dimension of aligned image-text embeddings, and D_e the dimension of normalized item embeddings used by the symbolic decoders. Text provides the BoW vector $\mathbf{x}_i^{(w)} \in \mathbb{R}_{\geq 0}^V$ [37], and speech provides the BoAW vector $\mathbf{x}_i^{(a)} \in \mathbb{R}_{\geq 0}^V$. These two count vectors form the reconstructable symbolic evidence from which topic descriptors are read.

For corpora with images, normalized aligned image-text embeddings $\mathbf{c}_i^{(t)}, \mathbf{c}_i^{(v)} \in \mathbb{R}^{D_c}$ from CLIP [20] are mapped by $g_{\boldsymbol{\psi}}(\cdot)$ into contextual evidence $\mathbf{u}_i \in \mathbb{R}^{2V}$ for the encoder. The model later reconstructs the original aligned embeddings $\mathbf{c}_i^{(t)}$ and $\mathbf{c}_i^{(v)}$, rather than $\mathbf{u}_i$. This auxiliary dense reconstruction retains contextual structure without adding another topic-descriptor layer.

3.3 Shared Hyperspherical Topic Inference

Because the encoder must combine sparse symbolic counts with normalized dense context, we infer the shared document latent direction directly on the unit sphere rather than in an unconstrained Euclidean latent space. This keeps document-specific semantics in direction and uses concentration to represent how confidently the observed modalities agree. The encoder uses block fusion, combining evidence blocks without assuming that dense features and count coordinates share the same coordinate semantics. Let $\mathbf{s}_i = \mathbf{x}_i^{(w)}$ when speech is absent, and let $\mathbf{s}_i = [\mathbf{x}_i^{(w)}; \mathbf{x}_i^{(a)}]$ when speech is observed. For corpora with images, $\mathbf{u}_i = g_{\boldsymbol{\psi}}(\mathbf{c}_i^{(t)}, \mathbf{c}_i^{(v)})$ is added as a separate contextual block. We write the evidence input as an ordered block tuple: $\mathcal{E}_i = (\mathbf{s}_i)$ without images and $\mathcal{E}_i = (\mathbf{s}_i, \mathbf{u}_i)$ with images, so every observed modality constrains the same posterior.

Given these evidence blocks, the shared encoder predicts a valid vMF posterior [33, 34] by normalizing the mean direction and keeping the concentration positive:

$$\begin{aligned} \mathbf{r}_i &= f_\phi(\mathcal{E}_i), \\ \boldsymbol{\mu}_i &= \frac{W_\mu \mathbf{r}_i}{\|W_\mu \mathbf{r}_i\|_2}, \\ \kappa_i &= \mathrm{softplus}(\mathbf{w}_\kappa^\top \mathbf{r}_i + b_\kappa) + \epsilon, \\ q_\phi(\mathbf{z}_i \mid \mathcal{E}_i) &= \mathrm{vMF}(\boldsymbol{\mu}_i, \kappa_i), \qquad \mathbf{z}_i \in \mathbb{S}^{K-1}. \end{aligned} \tag{3}$$

The mean direction $\boldsymbol{\mu}_i$ represents the document's semantic direction in topic space, and κ_i controls the posterior concentration. Evidence can therefore move this direction and adjust its certainty, while the sampled document latent direction remains on $\mathbb{S}^{K-1}$. We use the uniform distribution on $\mathbb{S}^{K-1}$ as the prior, which is equivalent to a zero-concentration vMF distribution. During training, $\mathbf{z}_i$ is sampled with the standard reparameterized vMF sampler [33, 34]; Appendix gives the sampling and KL details. Encoder inference and symbolic decoding therefore use the same directional geometry.

The shared document latent direction $\mathbf{z}_i$ is not a simplex-valued topic mixture. We define $\boldsymbol{\theta}_i = \mathrm{softmax}(\tau \mathbf{z}_i) \in \Delta^{K-1}$, where τ controls how strongly directional differences are converted into topic assignments. The decoders then use $\boldsymbol{\theta}_i$ as the document-topic mixture.

3.4 vMF Mixture Decoding for Symbolic Evidence

The decoder is where the topic representation becomes readable. A point topic vector [8, 9] can indicate the direction of a topic, but it cannot express how broad the topic should be. We therefore model each symbolic topic as a vMF component [36, 27, 29] in the corresponding normalized item space. For modality $m \in \{w, a\}$, let $\mathbf{E}^{(w)} = \{\mathbf{e}_j^{(w)}\}_{j=1}^{V} \subset \mathbb{S}^{D_e-1}$ and $\mathbf{E}^{(a)} = \{\mathbf{e}_j^{(a)}\}_{j=1}^{V} \subset \mathbb{S}^{D_e-1}$. Topic k has a mean direction $\boldsymbol{\eta}_k^{(m)}$ and concentration $\rho_k^{(m)}$.

The symbolic decoder scores each item, normalizes the scores, and mixes the resulting topic-level item distributions. The score is given by the vMF log-density up to a normalizing constant:

$$
\begin{aligned}
s_{kj}^{(m)} &= \log C_{D_e}(\rho_k^{(m)}) + \rho_k^{(m)} {\boldsymbol{\eta}_k^{(m)}}^{\top} \mathbf{e}_j^{(m)}, \\
\beta_{kj}^{(m)} &= \frac{\exp(s_{kj}^{(m)})}{\sum_{q=1}^{V} \exp(s_{kq}^{(m)})}, \\
\hat{x}_{ij}^{(m)} &= \sum_{k=1}^{K} \theta_{ik} \beta_{kj}^{(m)}.
\end{aligned}
\tag{4}
$$

For $m = w$, this gives topic-word distributions; for $m = a$, topic-acoustic-codeword distributions. In speech-text settings, the same $\boldsymbol{\theta}_i$ explains both modalities, coupling speech and text at the topic-mixture level. The concentration $\rho_k^{(m)}$ controls semantic granularity: larger values select a tight neighborhood around the topic direction, while smaller values allow a broader region.

3.5 Contextual Semantic Reconstruction

Symbolic reconstruction keeps topics interpretable, but sparse counts cannot capture the full image-text context. For corpora with images, we reconstruct the original aligned dense embeddings from the same document latent direction, where $\hat{\mathbf{c}}_i^{(t)} = B_t(\mathbf{z}_i)$ and $\hat{\mathbf{c}}_i^{(v)} = B_v(\mathbf{z}_i)$. These reconstruction heads are auxiliary: they encourage $\mathbf{z}_i$ to retain visual and contextual information, while the descriptor layer remains tied to BoW and BoAW reconstruction.

3.6 Training Objective

The training loss follows the same evidence roles as the architecture. For the full image-speech-text model, we minimize

$$
\mathcal{J} = \sum_{i=1}^{N} \left[\mathcal{L}_{\mathrm{KL}}(i) + \sum_{m \in \{w,a\}} \mathcal{L}_{\mathrm{sym}}^{(m)}(i) + \sum_{m \in \{t,v\}} \mathcal{L}_{\mathrm{ctx}}^{(m)}(i) \right]. \tag{5}
$$

The image-text and speech-text instances use the same objective after removing the loss terms for unobserved evidence. Missing encoder inputs are omitted in the same way. The KL term [38, 39] is $\mathcal{L}_{\mathrm{KL}}(i) = \mathrm{KL}\left(q_\phi(\mathbf{z}_i \mid \mathcal{E}_i) \,\|\, p(\mathbf{z}_i)\right)$, which regularizes the posterior toward the uniform hyperspherical prior.

For count evidence, we use a normalized negative log-likelihood, since BoW and BoAW are histograms over symbolic units:

$$\mathcal{L}_{\mathrm{sym}}^{(m)}(i) = -\frac{1}{\|\mathbf{x}_i^{(m)}\|_1} \sum_{j=1}^{V} x_{ij}^{(m)} \log \hat{x}_{ij}^{(m)}, \qquad m \in \{w, a\}. \tag{6}$$

For contextual evidence, we use cosine reconstruction, since aligned image-text semantics are directional:

$$\mathcal{L}_{\mathrm{ctx}}^{(m)}(i) = 1 - \cos(\mathbf{c}_i^{(m)}, \hat{\mathbf{c}}_i^{(m)}), \qquad m \in \{t, v\}. \tag{7}$$

The symbolic terms define the readable topic interface. The contextual terms retain dense semantics in the shared document latent direction, but they do not define topic descriptors. Different corpus settings only remove missing evidence blocks and their losses; the shared vMF posterior and the vMF-mixture symbolic decoder family stay unchanged.

4 Experiments

We organize the evaluation around three design claims: **(C1)** a shared hyperspherical document latent direction should improve the topic-set trade-off between coherence and diversity across modality regimes; **(C2)** BoAW should let speech participate as reconstructable count evidence rather than only as transcript text or an opaque dense feature; and **(C3)** image and speech should provide complementary evidence when aligned with text in the same corpus.

4.1 Experimental Setup

Datasets, metrics, and implementation. We evaluate five multimodal corpora, with dataset statistics shown in Table 1. The datasets cover the three settings addressed by our main claims: Twitter100k [40] and Flickr30k [41] for image-text modeling, LibriSpeech [42] and TEDLIUM-R3 [43] for speech-text modeling, and SpokenCOCO-Tri for aligned image-speech-text triples built from MS COCO 2014 and SpokenCOCO [44, 45]. Following standard topic modeling practice, we report coherence (C_v), Topic Diversity (TD), and Quality [46]. Quality is computed as the product of C_v and TD, so it reflects both semantic consistency and non-redundancy [9]. All metrics use the top 10 words of each topic, with $K \in \{10, 20, 50\}$. In result tables, red and blue mark column-wise best and second-best values. All experiments used a single NVIDIA RTX 4090 GPU with 24 GB VRAM, CUDA 13.0, and PyTorch 2.9.0. Preprocessing, implementation details, and sensitivity analysis for the concentration parameter are provided in the appendix.

Table 1: Dataset statistics.

Dataset	Modality	# Documents	Avg. length	Vocab size
Twitter100k	image-text	100,000	7.40	8,092
Flickr30k	image-text	31,014	17.98	2,174
LibriSpeech	speech-text	5,832	1648.74	17,095
TEDLIUM-R3	speech-text	2,370	1951.13	8,441
SpokenCOCO-Tri	image-speech-text	82,783	10.45	2,065

Baselines. We compare M3-HNTM with representative classical, neural, hyperspherical, and multimodal topic models. The text and geometry baselines include LDA [1], NeuralLDA [7], ETM [9], ZeroShotTM [11], CombinedTM [10], BERTopic [32], vONT [36], NVGMTM [31], and S2WTM [28]. The main multimodal baselines are Multimodal-ZeroShotTM, Multimodal-Contrast [14], and CEMTM [17]. We evaluate each method only on datasets that provide its required modalities. The external baselines provide a broad performance reference. On SpokenCOCO-Tri, we also compare against our image-text and speech-text variants, which give matched internal tests of whether the full tri-modal model draws useful signal from more than a single dominant branch.

Table 2: Topic modeling performance on Twitter100k and Flickr30k under different topic numbers.

Model	*Twitter100k*						*Flickr30k*					
	$K = 10$			$K = 20$			$K = 10$			$K = 20$		
	C_v	TD	Quality	C_v	TD	Quality	C_v	TD	Quality	C_v	TD	Quality
LDA (2003)	0.372	0.880	0.327	0.307	0.983	0.302	0.459	0.554	0.254	0.411	0.614	0.253
NeuralLDA (2017)	0.390	1.000	0.390	0.341	1.000	0.341	0.320	0.999	0.320	0.287	0.957	0.275
ETM (2020)	0.351	0.210	0.074	0.367	0.180	0.066	0.369	0.185	0.068	0.355	0.136	0.048
ZeroShotTM (2021)	0.398	0.926	0.369	0.404	0.663	0.268	0.444	0.875	0.388	0.472	0.617	0.291
CombinedTM (2021)	0.392	0.937	0.367	0.421	0.665	0.280	0.488	0.777	0.379	0.490	0.628	0.308
BERTopic (2022)	0.405	0.910	0.369	0.489	0.924	0.452	0.442	0.684	0.302	0.515	0.703	0.362
vONT (2023)	0.409	0.963	0.394	0.322	1.000	0.322	0.365	0.945	0.345	0.343	0.957	0.328
NVGMTM (2023)	0.385	0.686	0.264	0.325	0.506	0.164	0.282	0.686	0.193	0.318	0.489	0.156
Multimodal-ZeroShotTM (2024)	0.381	0.961	0.366	0.389	0.801	0.311	0.447	0.872	0.390	0.458	0.696	0.319
Multimodal-Contrast (2024)	0.501	0.838	0.420	0.506	0.915	0.463	0.337	0.735	0.248	0.434	0.776	0.337
S2WTM (2025)	0.316	0.952	0.300	0.356	0.927	0.330	0.481	0.805	0.387	0.512	0.771	0.395
CEMTM (2025)	0.437	0.992	0.433	0.408	1.000	0.408	0.453	0.959	0.434	0.411	0.906	0.372
M3-HNTM (image-text)	0.507	1.000	0.507	0.513	1.000	0.513	0.442	1.000	0.442	0.463	0.965	0.447

4.2 Main Results

Image-text benchmarks. Table 2 tests whether dense visual context improves word-level topic discovery while topics remain text-described. The broader pattern is not automatic dominance by image features. Earlier image-text baselines often outperform single-modal neural topic models, but their gains are modest and uneven, and they do not consistently exceed geometry-aware topic models. The clearest positive result is on Twitter100k, where the text is short and noisy. M3-HNTM improves Quality over the strongest external baseline at both $K = 10$ and $K = 20$. CEMTM is the closest current multimodal baseline through its LVLM-centered contextual route. Its competitive but lower Quality suggests richer contextual encoders alone do not solve the topic modeling problem unless outputs connect to a readable topic interface and a suitable latent geometry. Flickr30k provides a useful boundary case. Some baselines achieve higher isolated coherence, but M3-HNTM obtains the best Quality by preserving much higher diversity. This supports the first design claim: the shared hyperspherical document latent direction is useful at the topic-set level, where topics should remain coherent without becoming repeated variants of one visual scene **(C1)**. The appendix shows the same Quality ordering at $K = 50$.

Table 3: Topic modeling performance on LibriSpeech and TEDLIUM-R3 across topic numbers.

Model	*LibriSpeech*						*TEDLIUM-R3*					
	$K = 10$			$K = 20$			$K = 10$			$K = 20$		
	C_v	TD	Quality	C_v	TD	Quality	C_v	TD	Quality	C_v	TD	Quality
LDA (2003)	0.372	0.262	0.098	0.367	0.187	0.069	0.344	0.233	0.080	0.335	0.151	0.050
NeuralLDA (2017)	0.465	0.980	0.456	0.302	0.940	0.284	0.365	1.000	0.365	0.362	0.967	0.350
ETM (2020)	0.364	0.584	0.213	0.243	0.231	0.056	0.328	0.558	0.183	0.268	0.294	0.079
ZeroShotTM (2021)	0.430	1.000	0.430	0.423	0.912	0.386	0.302	0.970	0.293	0.359	0.759	0.272
CombinedTM (2021)	0.403	1.000	0.403	0.435	0.909	0.396	0.343	0.984	0.337	0.350	0.790	0.277
BERTopic (2022)	0.387	0.447	0.173	0.420	0.369	0.155	0.343	0.391	0.134	0.404	0.366	0.148
vONT (2023)	0.380	1.000	0.380	0.440	0.968	0.425	0.333	1.000	0.333	0.332	0.974	0.323
NVGMTM (2023)	0.374	1.000	0.374	0.395	0.939	0.371	0.528	0.157	0.083	0.566	0.122	0.069
S2WTM (2025)	0.362	0.978	0.354	0.429	0.907	0.389	0.414	0.912	0.378	0.401	0.830	0.333
M3-HNTM (speech-text)	0.523	1.000	0.523	0.455	1.000	0.455	0.411	1.000	0.411	0.431	1.000	0.431

Speech-text benchmarks. Table 3 tests whether speech provides topic evidence without replacing lexical interpretability. Across both speech-text corpora, M3-HNTM obtains the best Quality at $K = 10$ and $K = 20$, and the appendix shows the same order at $K = 50$. The point is not that acoustic codewords become human-readable topics. Instead, BoAW constrains the same topic mixture through reconstructible acoustic counts, while learned topics are still evaluated using word descriptors **(C2)**. TEDLIUM-R3 makes the metric trade-off clear. NVGMTM achieves the highest isolated C_v, but diversity collapses, producing very low Quality. M3-HNTM instead accepts slightly lower coherence here while keeping near-maximal TD and the best Quality. This is the intended behavior of shared directional inference: speech should organize a non-redundant topic set, rather than sharpen a small set of repeated descriptors **(C1)**.

Controlled tri-modal comparison. Table 4 provides a controlled test of tri-modal complementarity, since all M3-HNTM variants are evaluated on the same aligned corpus. The full model obtains the best Quality at every topic count, so the gain is not tied to a single topic granularity. The two

Table 4: Topic modeling performance on SpokenCOCO-Tri under different topic numbers.

	Num. topics	10			20			50		
	Metrics	C_v	TD	Quality	C_v	TD	Quality	C_v	TD	Quality
SpokenCOCO-Tri	LDA (2003)	0.410	0.878	0.360	0.435	0.811	0.353	0.370	0.760	0.281
	NeuralLDA (2017)	0.324	1.000	0.324	0.355	1.000	0.355	0.363	0.841	0.305
	ETM (2020)	0.239	0.121	0.029	0.236	0.083	0.020	0.226	0.063	0.014
	ZeroShotTM (2021)	0.371	0.920	0.341	0.450	0.699	0.314	0.373	0.313	0.117
	CombinedTM (2021)	0.371	0.994	0.369	0.401	0.794	0.319	0.389	0.395	0.154
	BERTopic (2022)	0.446	0.839	0.374	0.474	0.771	0.365	0.487	0.679	0.330
	vONT (2023)	0.354	1.000	0.354	0.354	0.977	0.346	0.355	0.950	0.337
	NVGMTM (2023)	0.257	0.721	0.185	0.285	0.557	0.159	0.302	0.343	0.104
	Multimodal-ZeroShotTM (2024)	0.326	0.963	0.314	0.428	0.788	0.337	0.385	0.364	0.140
	Multimodal-Contrast (2024)	0.439	0.801	0.352	0.410	0.928	0.381	0.437	0.849	0.371
	S2WTM (2025)	0.450	0.976	0.439	0.474	0.957	0.454	0.444	0.765	0.340
	CEMTM (2025)	0.496	0.904	0.448	0.408	0.910	0.371	0.437	0.886	0.387
	M3-HNTM (image-text)	0.469	0.995	0.467	0.482	0.990	0.477	0.514	0.820	0.427
	M3-HNTM (speech-text)	0.515	0.935	0.481	0.507	0.896	0.455	0.466	0.780	0.363
	M3-HNTM (full)	0.556	1.000	0.556	0.549	0.949	0.521	0.549	0.821	0.451

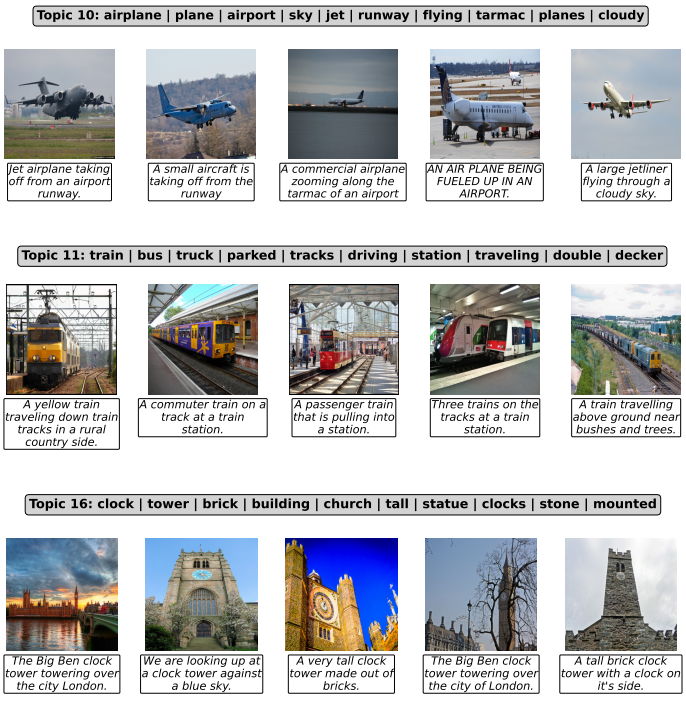
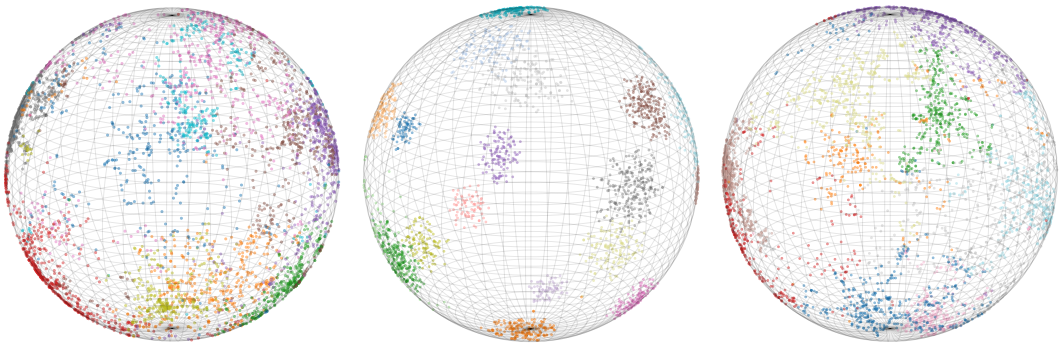


Figure 3: Representative word-level topics learned by the M3-HNTM on SpokenCOCO-Tri.

dual-modal branches also cross over: speech-text is better at $K = 10$, while image-text is better at $K = 20$ and $K = 50$. If one branch dominated, the full result could be interpreted as inheriting that branch. The cross-over instead gives a clearer explanation in terms of complementary evidence **(C3)**. The external baselines show that this is not simply a case where a multimodal encoder beats text-only modeling. On SpokenCOCO-Tri, earlier image-text multimodal baselines yield lower average Quality than S2WTM, so visual context alone does not outperform a strong geometry-aware topic model. CEMTM is the stronger current image-text baseline, but still does not use acoustic evidence and therefore cannot test the speech side of the tri-modal claim. The full M3-HNTM result is therefore not just a stronger visual-context result. Dense image-text evidence helps resolve visual context, BoW/BoAW count evidence anchors reconstructible descriptors, and the shared document latent direction requires both views to be fused before word-level topics are evaluated. Table 4 therefore supports complementarity on SpokenCOCO-Tri, with the limitation that it is a matched empirical comparison rather than a fully factorized causal proof of each modality's isolated effect **(C2, C3)**.



(a) Flickr30k (image-text). (b) TEDLIUM-R3 (speech-text). (c) SpokenCOCO-Tri (tri-modal).

Figure 4: t-SNE views of the learned document representations across image-text, speech-text, and tri-modal settings.

Qualitative topic structure. Figure 3 checks whether quantitative gains retain an interpretable topic interface. Selected topics pair readable word descriptors with retrieved image-text examples, such as station and vehicle scenes for the train topic and runway scenes for the airplane topic. This is the target qualitative pattern: dense visual and acoustic evidence can shape the posterior, while the final topic remains a word-level object that readers can inspect **(C2)**. Figure 4 provides the document-level version of Flickr30k, TEDLIUM-R3, and SpokenCOCO-Tri across the image-text, speech-text, and tri-modal settings. The visible clusters are consistent with a shared hyperspherical document latent direction that organizes examples by directional semantics **(C1)**. We treat Figure 4c as supporting evidence, not a substitute for the matched tri-modal comparison in Table 4. Additional SpokenCOCO-Tri topic examples appear in the appendix.

4.3 Ablation Study

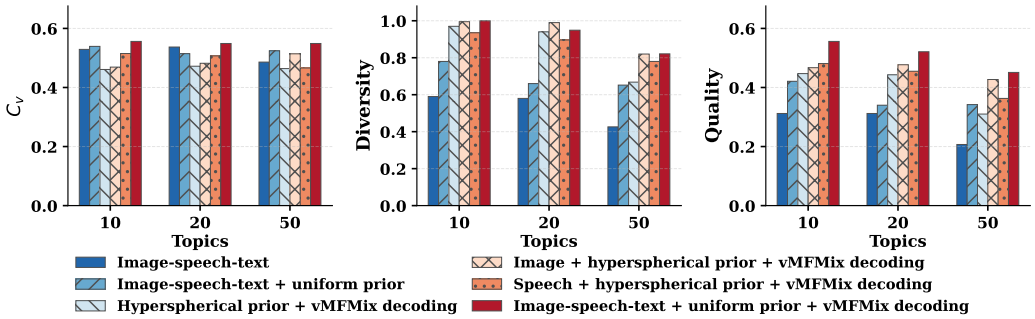


Figure 5: Staged ablation over model components and observed evidence on SpokenCOCO-Tri.

Figure 5 provides a staged check of the mechanism, rather than a fully isolated single-factor decomposition. The progression highlights three roles. The uniform prior improves the geometry of posterior inference, the vMF-mixture decoder preserves a symbolic descriptor path, and observed image or speech evidence sharpens the topic boundary beyond text alone. The full tri-modal model achieves the highest average Quality, but the main point lies in the overall progression: no single row accounts for the final result. The conclusion is therefore joint. Geometry alone does not recover the full model, symbolic decoding without added modalities remains below the dual-modal variants, and each dual-modal branch lacks evidence provided by the other. This is the evidence pattern targeted by M3-HNTM: different count-based and dense signals meet in a shared hyperspherical document latent direction, while the reported topics remain readable at the word level **(C1, C2, C3)**.

Table 5 complements the numerical ablation through one readable case where the topic boundary narrows. The text-only topic mixes animals, vehicles, and generic scene terms. Image context reduces this ambiguity, but the topic still spans livestock and pasture scenes. The full tri-modal model

Table 5: A representative 10-topic case from SpokenCOCO-Tri.

Variant	Representative topic words
M3-HNTM (Text only)	couple, horse, motorcycle, one, back, rid, elephant, behind, another, park
M3-HNTM (Image-text)	horse, horses, sheep, cow, cows, cattle, grazing, pasture, carriage, animals
M3-HNTM (Full)	horse, horses, carriage, drawn, pulling, cart, wagon, leading, barn, farm

produces a carriage-centered topic with words such as "drawn", "pulling", "cart", and "wagon". This change goes beyond replacing the word list. The added modalities narrow a broad co-occurrence pattern into a clearer scene concept, while keeping the descriptor symbolic and readable **(C2, C3)**.

4.4 Computational Cost

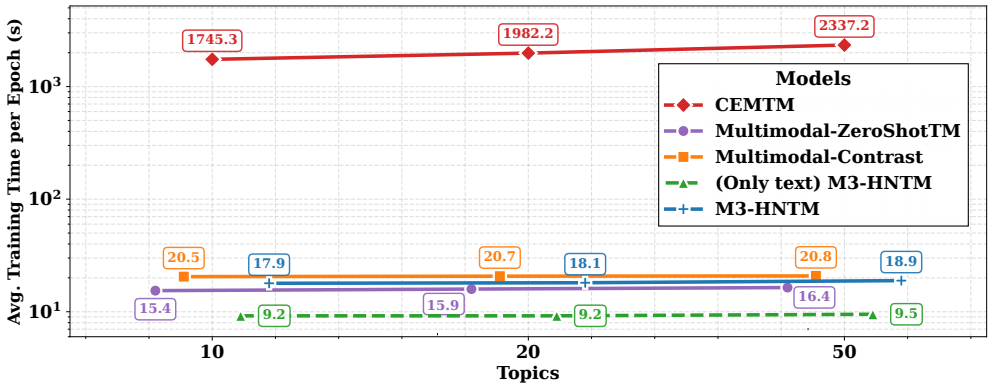


Figure 6: Training time per epoch on SpokenCOCO-Tri.

Computational cost is important because a natural route in multimodal topic modeling is to use an LVLM-centered contextual pipeline and let rich embeddings carry much of the modeling load. CEMTM is the strongest representative of this route in our comparison, but Figure 6 shows that it trains at a much higher cost on SpokenCOCO-Tri, taking 1745.3–2337.2 s per epoch. The full M3-HNTM model trains in 17.9–18.9 s per epoch, close to lightweight multimodal baselines such as Multimodal-ZeroShotTM (15.4–16.4 s) and Multimodal-Contrast (20.5–20.8 s). The claim is therefore not that M3-HNTM is the fastest model, but that its Quality gains come from geometry-matched evidence design rather than orders-of-magnitude heavier contextual encoding.

5 Conclusion

This paper treats multimodal topic modeling as evidence-role assignment: symbolic counts preserve topic interpretability, while dense aligned features improve posterior inference. M3-HNTM implements this view with a shared vMF posterior and vMF-mixture decoders for symbolic observations, placing heterogeneous evidence and concentration-aware topic descriptors in a common directional geometry. Across image-text, speech-text, and tri-modal benchmarks, M3-HNTM yields higher Quality under standard top-word metrics. Results on aligned tri-modal data further show that visual and acoustic evidence provide complementary topic signals.

Limitations and Future Work. The current study is bounded by English datasets, fixed topic numbers, and automatic coherence-diversity evaluation. Future work can extend the same evidence-role design to multilingual corpora, adaptive topic structures, richer acoustic context, and downstream corpus-analysis tasks.

Acknowledgments

The completion of this work was supported in part by the National Natural Science Foundation of China (62276106), Guangdong Basic and Applied Basic Research Foundation (2024A1515011767), the Guangdong Provincial Key Laboratory IRADS (2022B1212010006), and BNBU Start-up Research Fund (R0700171-27).

References

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

[2] Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. Neural topic modeling on hyperspheres: Spherical representation learning with von mises–fisher mixtures. *Neural Networks*, 200:108792, 2026.

[3] Zhiwen Luo, Manar Amayri, Wentao Fan, Koffi Eddy Ihou, and Nizar Bouguila. Parallel inference for cross-collection latent generalized dirichlet allocation model and applications. *Expert Systems with Applications*, 238:121720, 2024.

[4] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: a survey. In *International Joint Conference on Artificial Intelligence 2021*, pages 4713–4720. Association for the Advancement of Artificial Intelligence (AAAI), 2021.

[5] Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Flexible priors for cross-collection topic modeling: an approach with generalized dirichlet and beta-liouville distributions. *International Journal of Data Science and Analytics*, 22(1):83, 2026.

[6] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR, 2016.

[7] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[8] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *International conference on machine learning*, pages 2410–2419. PMLR, 2017.

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

[10] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th annual meeting of the ACL and the 11th IJCNLP*, pages 759–766, 2021.

[11] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 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, 2021.

[12] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18, 2024.

[13] Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. Msw-ntm: A spherical wasserstein autoencoder for multimodal neural topic modeling with llm-guided topic refinement. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3759–3763, 2026.

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

[15] Dejan Porjazovski, Tamás Grósz, and Mikko Kurimo. Topic identification for spontaneous speech: Enriching audio features with embedded linguistic information. In *2023 31st European Signal Processing Conference*, pages 396–400, 2023.

[16] Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. Multimodal topic discovery in web media via von mises-fisher mixture neural topic models. In *Proceedings of the ACM Web Conference 2026*, pages 8749–8752, 2026.

[17] Amirhossein Abaskohi, Raymond Li, Chuyuan Li, Shafiq Joty, and Giuseppe Carenini. Cemtm: Contextual embedding-based multimodal topic modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11686–11703, 2025.

[18] Gábor Gosztolya and Róbert Busa-Fekete. Ensemble bag-of-audio-words representation improves paralinguistic classification accuracy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:477–488, 2020.

[19] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

[21] Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Hyperspherical representation learning of axial data via axial vaes with watson distribution. *ACM Transactions on Knowledge Discovery from Data*, 20(4):1–21, 2026.

[22] Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Adaptive deep clustering via disentangled hyperspherical vaes with contrastive geometry optimization. *Neurocomputing*, page 132625, 2026. ISSN 0925-2312.

[23] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, August 2016.

[24] Xue Han, Zhiwen Luo, Zhixiang Li, Nizar Bouguila, Weifeng Su, and Wentao Fan. Hyperxrec: Unifying preference clusters and llm experts for robust explainable recommendations. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pages 2653–2661. International Joint Conferences on Artificial Intelligence Organization, 8 2026.

[25] Zhiwen Luo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Dynamic deep clustering of high-dimensional directional data via hyperspherical embeddings with bayesian nonparametric mixtures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 938–949, 2025.

[26] Zhixiang Li, Zhiwen Luo, Nizar Bouguila, Weifeng Su, and Wentao Fan. Disentangled representation learning for multi-view clustering via von mises-fisher hyperspherical embedding. *Neural Networks*, page 107802, 2025.

[27] Hafsa Ennajari, Nizar Bouguila, and Jamal Bentahar. Correlated topic modeling for short texts in spherical embedding spaces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(06):4567–4578, 2025.

[28] Suman Adhya and Debarshi Kumar Sanyal. S2wtm: Spherical sliced-wasserstein autoencoder for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23211–23225, 2025.

[29] Rui Wang, Yanan Wang, Xing Liu, Haiping Huang, and Guozi Sun. Bridging spherical mixture distributions and word semantic knowledge for neural topic modeling. *Expert Systems with Applications*, 256:124850, 2024.

[30] Dayu Guo, Zhiwen Luo, Nizar Bouguila, and Wentao Fan. Biovmnvtm: A geometry-aware neural topic model for biomedical text analysis via von mises–fisher mixtures. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3648–3653. IEEE, 2025.

[31] Yi-Kun Tang, Heyan Huang, Xuewen Shi, and Xian-Ling Mao. Neural variational gaussian mixture topic model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–18, 2023.

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

[33] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[34] Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4503–4513, 2018.

[35] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. Spherical topic models. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 903–910, 2010.

[36] Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. vontss: vmf based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457, 2023.

[37] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52, 2010.

[38] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[39] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

[40] Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938, 2017.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

[42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing*, pages 5206–5210, 2015.

[43] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208, 2018.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[45] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5284–5300. Association for Computational Linguistics, 2021.

[46] Zhiwen Luo, Manar Amayri, Wentao Fan, and Nizar Bouguila. Cross-collection latent beta-liouville allocation model training with privacy protection and applications. *Applied Intelligence*, 53(14):17824–17848, 2023.

[47] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state and accurately reflect our novel contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We include a discussion of the potential limitations of our work in the conclusions section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper does not present theoretical results; mathematical details are provided for model specification.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper has fully disclosed all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: We plan to make the code repository publicly available via GitHub after the paper is published, with detailed usage instructions and reproduction guidelines. The supplementary materials include core code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: We specify data splits, hyperparameters, optimizers, hardware details in the Experimental Setup section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: This paper does not include error bars or formal statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper provides sufficient information on the computer resource for experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We confirm that our research adheres fully to the NeurIPS Code of Ethics guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No].

    Justification: The paper does not specifically discuss the societal impacts of the research. This work is fundamental research focused on improving multimodal topic modeling, primarily for scientific and academic applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: We determine that our work poses no high-risk assets requiring special release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We credit all reused code and datasets with explicit citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not include crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not include human-subject experiments requiring IRB approval or participant risk disclosure.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method development in the paper does not involve LLMs in any important, original, or non-standard way.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Additional Method Details

A.1 Sampling from the Hyperspherical Posterior

In the main paper, we model the document posterior as

$$q_\phi(\mathbf{z}_i \mid \mathcal{E}_i) = \mathrm{vMF}(\boldsymbol{\mu}_i, \kappa_i),$$

where $\mathcal{E}_i$ denotes the available evidence blocks and $\mathbf{z}_i$ lies on the unit hypersphere. The sampling procedure follows the standard construction used in hyperspherical VAEs [33, 34]. First, an axial component r_i is sampled according to the concentration parameter κ_i. For $K = 3$, inverse-transform sampling admits a closed form, whereas higher-dimensional cases can be handled by rejection sampling. Second, an orthogonal unit vector $\mathbf{o}_i$ is obtained by normalizing a standard Gaussian sample. These two terms define the intermediate vector

$$\mathbf{z}_i' = \begin{bmatrix} r_i \\ \sqrt{1 - r_i^2}\, \mathbf{o}_i \end{bmatrix},$$

which is then rotated to the target mean direction $\boldsymbol{\mu}_i$ through a Householder transformation. This yields a differentiable sample on $\mathbb{S}^{K-1}$ while preserving the directional semantics encoded by the posterior.

A.2 KL Term for the vMF Posterior

The main objective uses

$$\mathrm{KL}(q_\phi(\mathbf{z}_i \mid \mathcal{E}_i) \,\|\, p(\mathbf{z}_i)),$$

where $p(\mathbf{z}_i)$ is the uniform distribution on $\mathbb{S}^{K-1}$. Since the uniform prior is the zero-concentration vMF distribution, the KL term depends on the posterior concentration and the vMF normalization constant. For the $d = K$ posterior used in the main paper, this normalizer is

$$C_K(\kappa) = \frac{\kappa^{K/2-1}}{(2\pi)^{K/2} I_{K/2-1}(\kappa)},$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind. The KL can be written as

$$\mathrm{KL}_i = \kappa_i \frac{I_{K/2}(\kappa_i)}{I_{K/2-1}(\kappa_i)} + \log C_K(\kappa_i) - \log\left(\frac{\Gamma(K/2)}{2\pi^{K/2}}\right).$$

This form clarifies the regularization effect behind the hyperspherical posterior. Under an isotropic Gaussian prior, the KL term penalizes both posterior mean and variance, which can pull document-specific latent codes toward a shared Euclidean origin. With a uniform prior on $\mathbb{S}^{K-1}$, the vMF–uniform KL is rotationally invariant and depends only on the posterior concentration κ_i, not on the mean direction $\boldsymbol{\mu}_i$ [33, 34]. The regularizer therefore discourages overly sharp posteriors through concentration while leaving directional semantics free of any preferred origin. This is the technical reason the main paper treats hyperspherical regularization as geometry-matched support for normalized semantic evidence [23, 22, 21].

This expression is used only as an implementation detail. The main paper keeps the KL term symbolic because the modeling point is the hyperspherical regularization, not the Bessel-function algebra.

B Additional Experimental Details

B.1 Evaluation Protocol

We compare models within the modality settings they can support. Twitter100k and Flickr30k instantiate the image-text setting, LibriSpeech and TEDLIUM-R3 instantiate the speech-text setting, and SpokenCOCO-Tri supports the two dual-modal variants as well as the full tri-modal model. This grouping avoids invalid cross-modality comparisons and keeps the appendix aligned with the three claims tested in the main paper: hyperspherical topic geometry, reconstructable speech evidence, and visual–acoustic complementarity.

Metric definitions. For each learned topic t_j, let $\mathcal{W}_j = \{w_{j1}, w_{j2}, \ldots, w_{jn}\}$ denote the top n words used for evaluation. In all reported experiments, $n = 10$. Following the C_v coherence construction [47], each top word w_{ji} is represented by a context vector $\boldsymbol{v}_{ji}$ derived from sliding-window co-occurrence statistics against the topic's top-word set. Let $\boldsymbol{v}_{cj} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_{ji}$ be the centroid of these context vectors. The topic-level coherence is

$$C_v(t_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{v}_{ji}^{\top}\boldsymbol{v}_{cj}}{\|\boldsymbol{v}_{ji}\|\,\|\boldsymbol{v}_{cj}\|}.$$

We report C_v as the mean of $C_v(t_j)$ over the K topics. Topic Diversity follows the standard top-word uniqueness score [9]:

$$\mathrm{TD} = \frac{\left|\bigcup_{j=1}^{K} \mathcal{W}_j\right|}{Kn}.$$

Thus C_v measures semantic consistency within each topic, TD measures lexical non-redundancy across the topic set, and Quality is $C_v \times \mathrm{TD}$.

B.2 Implementation Settings Inherited from the Base Models

Unless otherwise specified, all methods use the shared experimental protocol. Each dataset is split into 80% training and 20% test data with the same split across compared methods. Text is tokenized, lemmatized, and filtered with standard stopword removal. We remove words appearing in fewer than 10 documents and words appearing in more than 5% of all documents. Lexical item embeddings use *glove-wiki-gigaword-100* [48]. Speech frame embeddings are extracted with *wav2vec2-base-960h* [19]; the BoAW codebook is built with k-means, and its size is set to the BoW vocabulary size of the corresponding corpus. Aligned image-text contextual embeddings are extracted with *clip-ViT-B-32* [20]. Baselines use the public implementations listed below and follow their reported settings. An anonymized implementation is included in the supplementary material.

For our model, the encoder consists of two hidden layers with 100 units each and ReLU activations, followed by dropout with rate 0.2. The decoder uses vMF mixture reconstruction heads for textual BoW and acoustic BoAW distributions, and linear topic-feature reconstruction heads for image and text embeddings. We optimize the model with Adam using a learning rate of 0.002, batch size 256, and a maximum of 200 epochs. In the default setting, the learnable document-posterior concentration κ_i and the topic-decoder concentrations $\rho_k^{(m)}$ are initialized at 50. Figure 7 reports the sensitivity of these two concentration families. The softmax temperature is set to $\tau = 10$. Randomized training steps use seed 42 for reproducibility.

B.3 Baseline Implementations

Table 6 lists the public implementations used for the external baselines. The role column explains why each baseline is included in the comparison set, rather than repeating the related-work discussion from the main paper.

Table 6: Baseline implementation sources.

Baseline	Role in the comparison	Public implementation
LDA, NeuralLDA, ETM	Classical, neural, and embedding-aware text topic models	OCTIS models
ZeroShotTM, CombinedTM	Contextualized text topic models	Contextualized topic models
BERTopic	Clustering-based contextual topic model	BERTopic repository
vONT	Hyperspherical variational topic model	vONTSS repository
NVGMTM	Gaussian-mixture neural topic model	NVGMTM repository
Multimodal-ZeroShotTM, Multimodal-Contrast	Image-text multimodal topic models	Multimodal topic repository
S2WTM	Spherical sliced-Wasserstein topic model	S2WTM repository
CEMTM	LVLM-based image-text topic model	CEMTM repository

C Additional Quantitative Details

C.1 Additional 50-Topic Dual-Modal Results

Table 7: Image-text results at $K = 50$ on Twitter100k and Flickr30k.

Dataset	Model	C_v	TD	Quality
Twitter100k	LDA (2003)	0.264	0.988	0.261
	NeuralLDA (2017)	0.378	0.934	0.353
	ETM (2020)	0.375	0.171	0.064
	ZeroShotTM (2021)	0.436	0.473	0.206
	CombinedTM (2021)	0.451	0.520	0.234
	BERTopic (2022)	0.381	0.947	0.360
	vONT (2023)	0.339	1.000	0.339
	NVGMTM (2023)	0.345	0.369	0.127
	Multimodal-ZeroShotTM (2024)	0.430	0.499	0.214
	Multimodal-Contrast (2024)	0.507	0.837	0.424
	S2WTM (2025)	0.369	0.921	0.339
	CEMTM (2025)	0.397	0.916	0.364
	M3-HNTM (image-text)	0.595	0.997	0.593
Flickr30k	LDA (2003)	0.376	0.704	0.265
	NeuralLDA (2017)	0.324	0.706	0.229
	ETM (2020)	0.375	0.110	0.041
	ZeroShotTM (2021)	0.495	0.346	0.171
	CombinedTM (2021)	0.497	0.334	0.166
	BERTopic (2022)	0.500	0.746	0.373
	vONT (2023)	0.317	0.764	0.242
	NVGMTM (2023)	0.309	0.255	0.079
	Multimodal-ZeroShotTM (2024)	0.447	0.349	0.156
	Multimodal-Contrast (2024)	0.442	0.699	0.309
	S2WTM (2025)	0.508	0.631	0.321
	CEMTM (2025)	0.450	0.839	0.377
	M3-HNTM (image-text)	0.462	0.824	0.381

Table 8: Speech-text results at $K = 50$ on LibriSpeech and TEDLIUM-R3.

Dataset	Model	C_v	TD	Quality
LibriSpeech	LDA (2003)	0.382	0.184	0.070
	NeuralLDA (2017)	0.280	0.928	0.260
	ETM (2020)	0.220	0.092	0.020
	ZeroShotTM (2021)	0.386	0.625	0.241
	CombinedTM (2021)	0.383	0.583	0.223
	BERTopic (2022)	0.477	0.375	0.179
	vONT (2023)	0.412	0.983	0.405
	NVGMTM (2023)	0.326	0.818	0.267
	S2WTM (2025)	0.396	0.510	0.202
	M3-HNTM (speech-text)	0.451	0.988	0.446
TEDLIUM-R3	LDA (2003)	0.338	0.095	0.032
	NeuralLDA (2017)	0.375	0.922	0.346
	ETM (2020)	0.262	0.140	0.037
	ZeroShotTM (2021)	0.308	0.645	0.199
	CombinedTM (2021)	0.375	0.670	0.251
	BERTopic (2022)	0.462	0.432	0.200
	vONT (2023)	0.311	0.964	0.300
	NVGMTM (2023)	0.554	0.087	0.048
	S2WTM (2025)	0.362	0.630	0.228
	M3-HNTM (speech-text)	0.417	0.971	0.405

C.2 Sensitivity to Concentration Parameters

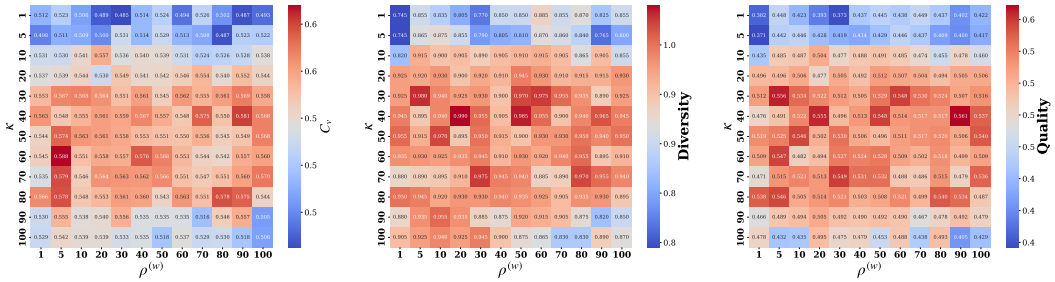


Figure 7: Sensitivity of the full model on SpokenCOCO-Tri to the initialization values of document-posterior and symbolic topic-decoder concentrations $(\kappa, \rho^{(w)})$.

M3-HNTM has two concentration families tied directly to the hyperspherical design. The document-posterior concentration κ controls how tightly the latent document direction concentrates around its mean. The symbolic topic-decoder concentration $\rho^{(w)}$ controls how narrow or broad each symbolic topic region is. Both denote initialization values and are updated during training.

Let $\mathcal{G} = \{1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. On SpokenCOCO-Tri, we sweep $(\kappa, \rho^{(w)}) \in \mathcal{G} \times \mathcal{G}$ and report C_v, topic diversity, and Quality in Figure 7. The best C_v is 0.588 at $(\kappa, \rho^{(w)}) = (60, 5)$, the best diversity is 0.990 at $(40, 20)$, and the best Quality is 0.561 at $(40, 90)$. The important pattern is the Quality heatmap, because Quality is the metric used throughout the paper to combine coherence and diversity. Very small document-posterior initialization gives weaker Quality: when $\kappa \in \{1, 5\}$, the best Quality is 0.449. In contrast, the broad region $\kappa \in [20, 80]$ and $\rho^{(w)} \in [10, 100]$ gives Quality between 0.477 and 0.561, with an average of 0.515. This supports the first design claim without relying on a single isolated hyperparameter pair.

D Additional Qualitative Topic Examples

D.1 Representative Topics Across Datasets

Table 9 pairs selected text-only and multimodal topics from the same dataset around a shared top-word anchor. The table is a qualitative companion to the quantitative results: added evidence should move a broad or mixed topic toward a sharper semantic boundary, while the topic remains readable as a word list.

Table 9: Paired representative topics with and without auxiliary modality evidence.

Setting	Topic words / boundary change
Twitter100k: politics	
Text only	trump, woman, top, medium, president, state, job, social, business, plan
Image-text	trump, clinton, hillaryclinton, hillary, president, obama, realdonaldtrump, donald, demsinphilly, potus
Change	Broad political and social-business words become a specific U.S. election event topic.
Twitter100k: beauty	
Text only	make, way, find, hair, beauty, keep, face, perfect, wear, makeup
Image-text	makeup, skin, beauty, hair, palette, products, lip, bbloggers, color, giveaway
Change	Generic action words are suppressed, leaving a concrete beauty-product topic.
LibriSpeech: public affairs	
Text only	administration, congress, holland, federal, senate, territory, election, tax, treaty, slavery
Speech-text	congress, administration, slavery, athenian, territory, federal, holland, election, senator, treaty
Change	The public-affairs theme remains readable while speech evidence stabilizes the register.
LibriSpeech: abstract discourse	
Text only	essence, product, economic, external, function, necessarily, duration, exists, complex, namely
Speech-text	essence, external, product, economic, function, necessarily, duration, exists, complex, scientific
Change	The same abstract discourse topic is preserved with a cleaner technical register.
SpokenCOCO-Tri: airplane	
Text only	large, clock, building, umbrella, sky, fly, airplane, plane, tower, tall
Tri-modal	airplane, plane, airport, runway, jet, sky, tarmac, passenger, planes, jetliner
Change	Mixed object and sky terms become a concrete airport scene.
SpokenCOCO-Tri: wildlife	
Text only	field, grass, walk, giraffe, tree, zebra, green, cow, bench, sheep
Tri-modal	grass, giraffe, field, elephant, walking, zebra, sheep, tree, herd, elephants
Change	Loose animal and background words become a coherent herd scene.

These paired cases illustrate the same pattern behind the main quantitative trend. Twitter100k shows visual context sharpening short, noisy posts; LibriSpeech shows speech evidence constraining the topic without removing word-level interpretability; SpokenCOCO-Tri shows boundary sharpening from broad visual co-occurrences toward concrete airport and wildlife-scene topics.

D.2 Additional SpokenCOCO-Tri Topic Sample

Figure 8 extends the three SpokenCOCO-Tri examples in Figure 3 using the same visualization protocol. These additional cases pair word-readable descriptors with image-caption examples of concrete scenes.

Topic 2: snow | surfboard | skis | wave | slope | ocean | skiing | ski | snowboard | snowy



A person riding a snowboard on a rail, in the snow.



A skier is weaving through the copious amount of snow.



Care of cross-country skis on a snowy mountain with trees.



A person is skiing down a steep mountain.



A skier arriving at the bottom of a ski slope

Topic 3: motorcycle | bike | motorcycles | bicycle | bikes | motor | police | bicycles | scooter | helmet



a side view of a vintage style harley davidson motorcycle



A closed parking structure packed full of motor scooters.



A motor cycle on display at a motorcycle event.



A lined up group of motor bikes on a road.



A motorcyclist rides properly dressed for the weather.

Topic 8: baseball | tennis | ball | player | court | bat | game | racquet | playing | racket



A woman hitting a tennis ball on a court.



a baseball player stretching while holding a baseball bat on a baseball field.



A woman that is on a tennis court with a racquet.



A baseball player is at stance to hit a thrown pitch.



A woman males a forehand shot during a tennis match.

Topic 9: laptop | computer | desk | keyboard | mouse | monitor | screen | computers | office | laptops



A desk topped with a desktop computer monitor and a laptop.



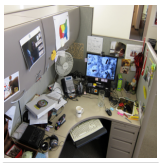
A laptop computer and a computer screen on display.



The back of a laptop has a sticker labeled "I Loves the 'Diff".



A work desk cluttered with stamps and work supplies



A cluttered cubicle with computer and coffee mug.

Figure 8: Additional word-readable topics and retrieved examples from the full M3-HNTM model on SpokenCOCO-Tri.