IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 February 2025, accepted 11 February 2025, date of publication 14 February 2025, date of current version 20 February 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3542562*

RESEARCH ARTICLE

# Explainable Video Topics for Content Taxonomy: A Multimodal Retrieval Approach to Industry-Compliant Contextual Advertising

**WARUNA DE SILVA, (Graduate Student Member, IEEE), AND ANIL FERNANDO, (Senior Member, IEEE)**
Department of Computer and Information Sciences, University of Strathclyde, G1 1XQ Glasgow, U.K.
Corresponding author: Anil Fernando (anil.fernando@strath.ac.uk)

**ABSTRACT** Owing to the increased video content consumption in recent years, the need for advanced contextual advertising methods that leverage increasing user engagement and relevance on advertisement-based video-on-demand platforms has increased. Traditional behavior-based advertisement targeting is waning, particularly owing to the recent strict privacy policies that favor user consent and privacy. This study proposes an innovative approach for integrating advanced natural language processing with multimodal analysis for video contextual advertising. To this end, transformer-based architectures, specifically BERTopic, computer vision techniques, and large language models were used to extract sets of topics from visual and textual video data automatically and systematically. The proposed framework decodes the taxonomy of content efficiently through videos in different levels of noise and languages. Empirical analysis of the YouTube-8M dataset shows the potential for the approach to change the paradigm in video advertising. Built to be scalable and easily adaptable, this solution can handle multifarious and complex user-generated content well, suited for a wide range of applications across various media platforms.

**INDEX TERMS** Natural language processing, video contextual advertisements, multimodal fusion, topic modeling, BERTopic, contextual taxonomy standards, multi-label classification.

## I. INTRODUCTION

Advertisement-based video-on-demand (AVoD) has emerged as a viable advertisement model owing to the significant increase in video consumption in recent years [1]. AVoD platforms earn revenue primarily through advertising via real-time bidding [2]. Several privacy-friendly advertisement measurement and targeting methods have been developed owing to the reduction in behavior-based advertisement targeting brought about by regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act [3]. These laws advocate for an opt-in consent-based model, which encourages consumers to move towards privacy-preserving technologies. This has heightened the requirement for scalable methods capable of mapping media assets to industry taxonomies while ensuring that compliance and monetization remain effective. Contextual advertising has emerged as a promising compromise. It operates via bets placed on a cookie-less uncluttered environment and delivers advertisements perfectly suited to the type of content being viewed based on semantic video indexing and deep learning processes [4]. Given that videos consist of rich multimodal data, such as audio, visual, and textual data, this is a new research paradigm for multimodal approaches to handle various data types [5], [6]. Furthermore, integrating support for multilingual enablement can extend reach, allowing users to execute advertising strategies targeting all relevant languages and regions, without alienating potential audiences [7].

Some studies have also suggested that thematically congruent advertisements improve both memorability and attitude toward advertisements in high-arousal contexts [8]. An in-

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

depth analysis was conducted to identify broader thematic elements that are capable of making targeted video classifiers more efficient at adding value and engaging viewers. This approach delivers higher memorability and positive viewer attitudes using explicit and implicit contexts, thereby enhancing the effectiveness of contextual advertising [9]. Beyond contextual advertising, topic modeling, which is a natural language processing (NLP) method, offers the possibility of inferring explicit and implicit themes [10]. These topics are aligned with the content taxonomy more precisely to enable finer targeting in a larger advertising ecosystem.

In this paper, we describe the utilization of transformer-based architectures, a family of models behind many recent advances in NLP and computer vision, to decode content taxonomies. Specifically, we employ distinct BERTopic models to extract and analyze topics from both image and textual data embeddings. These models systematically inferred topics based on visual modalities and multilingual transcripts. In addition, we leverage large language models (LLMs) to enhance the coherence and explainability of the topic representations. The inferred topics were systematically mapped to pertinent taxonomies, enabling nuanced and comprehensive analysis of multimodal video data. This technique effectively handles varying noise levels across video modalities and flexibly identifies content taxonomy based on outcomes across multiple video modalities. It exhibits both flexibility and scalability, particularly when dealing with user-generated content, which is often inconsistent, unstructured, and varies significantly in terms of quality.

The deployment of this framework is expected to increase the precision and engagement efficiency of advertisements significantly, effectively enabling programmatic advertisements on AVoD platforms. For this purpose, we performed an empirical analysis of a semi-automated processed sample dataset, YouTube-8M. Video content intrinsically involves multiple instances of a class occasionally; therefore, we used multi-label classifiers to address the complexity of video data comprising multiple types and categories. This study represents an advancement in the field of modeling contextual advertisements and video topic modeling.

The remainder of this paper is organized as follows. Related studies on contextual advertising and video topic modeling are discussed extensively in Section II. The proposed methodology, including the data sources and structure, is described in Section III. The experimental and analytical results are presented in Section IV. The application perspectives and directions of future research are discussed in Section V. Finally, the paper is concluded in Section VI by summarizing the key contributions and insights of this paper.

## II. RELATED WORK

The increasing demand for AVoD services is evident as major media players, such as Netflix, Rakuten, Discovery, Amazon, and Comcast, have already launched or plan to launch such services imminently [11], [12]. Personalized advertising approaches harness viewer data to display advertisements that are targeted and, therefore, not intrusive. This leverages viewers as influencers in the production of content and the targeting of audiences while monetizing viewer attention [13]. The GDPR and similar new regulations, such as the upcoming Digital Services Act, are expected to complicate the capture and use of personal data for targeted advertising by default [14]. One promising alternative is contextual advertising, which considers the context in which advertisement media are placed [4].

### A. CONTEXTUAL ADVERTISING

VideoSense [15] was introduced for contextual advertisements on video platforms. It utilizes elements such as titles, tags, queries, and local visual-aural features such as color, motion, and audio. Okada et al. [16] studied the process of selecting advertisements for placement on videos. They considered different forms of textual metadata created by users and stored them on a host webpage. These included titles, keywords, descriptions, categories, and comments, and were used to select relevant advertisements. This avoids the need to process images and videos elaborately. In Salad [17], a convolutional neural network (CNN) was adopted for feature extraction and selection of the most salient advertisements. It aligns text with visual content, preserving the context using high-level features obtained from a deep neural network for optimal relevance of advertisements in online videos. However, the aforementioned studies relied on metadata or a set of visual/aural cues without performing a proper analysis of the video content; therefore, they did not consider context-specific nuances that are essential for advertising. Similarly, Zhang et al. [18] conducted research on online video advertising to optimize the balance between advertisement intrusiveness and relevance. Their work incorporated a conventional histogram of oriented gradient features for generic object detection and deep CNNs for class-specific tasks, such as gender recognition in clothing retrieval. Moreover, they considered and recommended the incorporation of deep neural networks to enhance object detection performance. Wang et al. [19] combined hue, saturation, and value color histograms with Oriented FAST and Rotated BRIEF(ORB) features to offer users a highly detailed measure of content similarity and, therefore, focused on the overall relevance of a scene rather than single objects in the scene. AI advances allow for deeper semantic analysis, enhancing ad placement. There have been three contextual factors in ad acceptance that have been studied: applicability, affective tone, and consumer engagement [20]. An interesting step forward was made with the development of the multimodal approach, DEEP-AD [21], which utilizes a temporal video segmentation algorithm to relate advertisements with video content semantically. The algorithm segments videos into stories by analyzing the visual, audio, and semantic features using a battery of deep CNNs. In DEEP-AD, the semantic descriptions of both video scenes and advertisements are employed to ensure contextual relevance. Object and place recognition methods are applied

to derive these semantic descriptions, thereby enhancing the precision of ad placement. Another contextual ad platform, SemanticAd [22], extends a similar idea and aims at specific ad placement at semantic boundary positions in a video, not semantic compatibility alone. Unlike DEEP-AD's semantic deep-description direction, SemanticAd utilizes story unit extraction for ad and video segmentation and ad mapping in terms of visual, audio, and semantic discontinuity. Further, Song et al. [23] proposed a multimodal approach combining various types of CNNs to extract multimodal features and obtain a unified representation of over 140 movie video clips based on semantics, objects, scenes, sentiments, colors, and audio. They modeled topics using a semantics-based model which is based on an I3D framework. The I3D framework is suitable for the short shots used in recognition tasks and precise action recognition. However, it does not function effectively on long-range videos [24]. Existing video segmentation methods have predominantly targeted short videos characterized by clear visual changes and simple patterns [25]. These unique properties of short videos can be learned using supervised configurations, yielding models that are highly resistant to the longer and more subtle properties of long-form video materials. NLP techniques are relatively prevalent in traditional advertisement formats, including search, social, web, and classified advertisements [26]. Although a few studies considered the use of NLP in video advertising until 2022 [26], recently, with the growing importance of video content, an interest in the exploration of advanced NLP concepts from multiple perspectives has been observed [20], [27]; hence, deeper insights with innovative applications are expected in this domain. Finally, Explainable AI (XAI) is becoming a critical issue in digital marketing, in a direction to prevent a lack of transparency in AI-powered ad placement. In [28], a model incorporates CTR prediction, visual heatmaps, and LLM analysis for brands, with an objective towards providing increased transparency in ad targeting. In this work, ad effectiveness is stressed to be increased with interpretability in terms of audience engagement through XAI.

### B. VIDEO TOPIC MODELING

The NLP method of topic modeling was primarily developed for text analysis, using popular models such as Latent Dirichlet Allocation (LDA), which employs probabilistic approaches to discover hidden themes in large text corpora [29]. In textual data, a topic is represented as a "bag of words". For videos, a similar concept can be applied, with a "bag of features" representing the video content. From a semantic perspective, a "topic" in video analysis is represented by objects, behaviors, activities, events, etc. [30]. To adapt LDA to videos, features extracted from each frame may be quantized to the nearest visual words in a predefined dictionary. However, unlike language models, video analysis does not include predefined words. Therefore, a global embedding method such as word2vec [31] is not suitable, which leads to difficulties in semantic measurements. Although several approaches that combine language and vision have partially addressed this issue, designing effective information embedding methods for topic-based video analysis remains challenging [32]. Although BERTopic [33] was not designed specifically for video topics, its fundamentals can be applied to analyze key images in videos using image embeddings and audio transcripts in conjunction with sentence transformers, because sentence semantics can capture more powerful and context-rich information than individual words. Recent works [34] have shown that BERTopic outperforms other models, such as LDA [29], non-negative matrix factorization [35] and contextualized topic models [36], thus placing it as the strongest candidate for both multimodal and textual topic analysis. This robustness and semantic depth underpin its selection for video topic modeling in the present study.

## III. PROPOSED SYSTEM

Inspired by recent neural network-based unsupervised approaches to topic modeling, we propose a model that accepts, as its input, a video with 1) video frames and 2) audio transcripts as sentences to derive a representation of both visual and textual data. This method allows the model to identify and segment different topics in a video by defining boundaries for thematic content and content changes. In this method, all possible themes in a video are to be captured as topics, avoiding missing critical themes owing to the nature of the video.

**Given:** A video $X_i$ with transcript $A_i$ as a sequence of sentences $\{s_1, s_2, \ldots, s_n\}$ and video frames $V_i = \{k_1, k_2, \ldots, k_m\}$.

**Predict:** The set of taxonomies $\{\mu_1 T_1, \mu_2 T_2, \ldots, \mu_p T_p\}$. Each taxonomy is associated with a multiplier $\mu_i$ representing term frequency, i.e., the frequency and relevance of taxonomy based on the mapped topics.

This framework consists of five major components, as illustrated in Figure 1. The stages include 1. visual topic model training, 2. textual topic model training, 3. feature extraction, 4. video topic inference, and 5. industry content taxonomy association.

Both visual and textual topic models are trained using BERTopic [33], which leverages embeddings and hierarchical clustering to create coherent topic representations and yields robust and interpretable topic models. These key visual frames and audio transcripts are used as input in the video topic inference phase. Eventually, these topics are mapped to the IAB TechLab content taxonomy [37] to enable precise contextual advertising. This methodology provides a thematic understanding of video content for contextual relevance.

### A. TOPIC MODELING WITH BERTOPIC

Two separate BERTopic pipelines were developed for visual data and audio transcripts based on the modular architecture of BERTopic, which selects appropriate components to con-

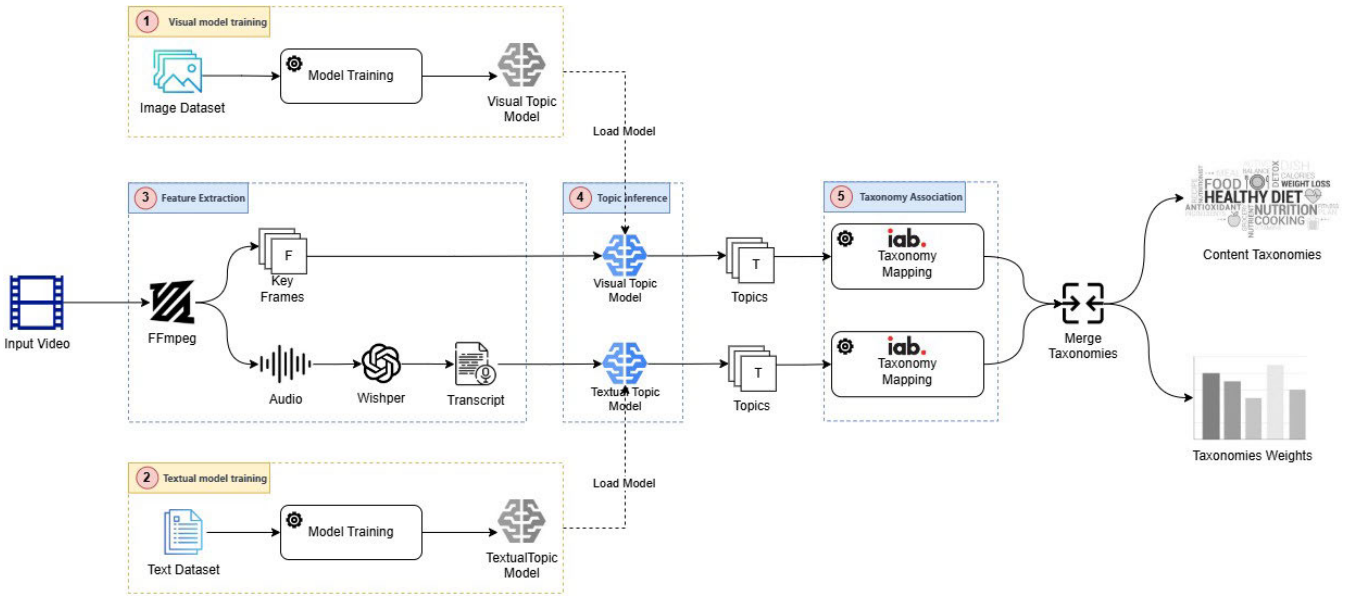


**FIGURE 1.** Multimodal embedding architecture for contextual video analysis.

front the challenges presented by each modality. BERTopic uses embeddings, dimensionality reduction, clustering, and topic-representation pipelines to generate coherent topics. Hyperparameter tuning is a significant requirement for mapping a model with the characteristics of each modality to achieve coherence topics from each stream.

For each video, $X_i$, we initially partition the data into visual and audio components, enabling the use of tailored models for each modality.

### 1) COMMON COMPONENTS OF BERTOPIC

- **Embedding**: Input data were transformed into rich feature sets using advanced embedding models. These embeddings capture the essential characteristics of the data, enabling effective subsequent processing.
- **Dimensionality Reduction**: The fine-tuning of the Uniform Manifold Approximation and Projection (UMAP) [38] hyperparameters during dimension reduction balances the simplicity of the data while preserving the complex structures corresponding to each modality.
- **Clustering**: Fine-tuned hyperparameters were provided for Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [39] to provide clear and meaningful visual clustering, facilitating the discovery of heterogeneous thematic content.
- **Topic Representation Formation**: Topic vectors $\{T_{V_{i1}}, T_{V_{i2}}, \ldots, T_{V_{in}}\}$ and $\{T_{A_{i1}}, T_{A_{i2}}, \ldots, T_{A_{in}}\}$ are created for visual and audio data, respectively. In BERTopic, to obtain an accurate representation of the topics from the bag-of-words matrix, the term frequency-inverse document frequency (TF-IDF) [40] is adjusted to work on the cluster level instead of the document level. This adjusted TF-IDF representation is called c-TF-IDF, and it considers the essential differences between documents in different clusters:

$$w_{x,c} = \mathrm{tf}_{x,c} \times \log\left(1 + \frac{A}{f_x}\right) \tag{1}$$

where:

$\mathrm{tf}_{x,c}$ = frequency of word $x$ in class $c$
$f_x$ = frequency of word $x$ across all classes
$A$ = average number of words per class

### 2) BERTOPIC(VISUAL DATA)

As depicted in Figure 2, the following steps are tailored for visual modality.

- **Embedding with CLIP-ViT** [41]: The "CLIP-ViT-B-32" embedding model is deployed to transform images into a rich feature set that is encoded using a Vision Transformer (ViT) with a base-size (B) architecture and 32 attention heads.
  Models such as the ViT [42], ResNet [43] and VGG-16 [44] tend to excel when it comes to visual features extracted by clustering. However, these traditional methods mostly fail [45] to deliver semantically meaningful clusters even while being quite effective in clustering visually similar images with approaches such as Nearest Neighbor Matching [46], [47]. CLIP tries to overcome this limitation by aligning visual and textual data in a common feature space, thus enabling clustering that captures both the visual and semantic relationships. CLIP has already outperformed over 20 state-of-the-art visual-model-based methods such as ResNet [45]. With ViT embeddings and semantic richness, CLIP-ViT ensures that clustering achieves both robustness and meaningful grouping [48]. This dual capability enhances

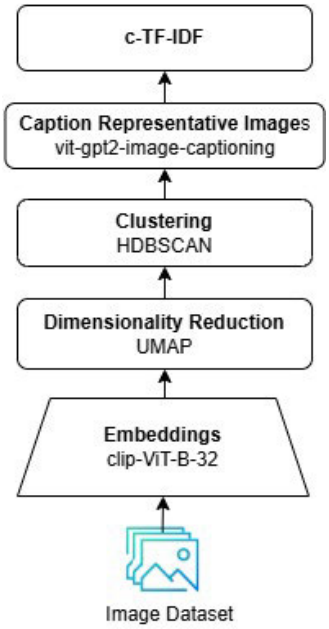


**FIGURE 2.** Visual data processing workflow using BERTopic.

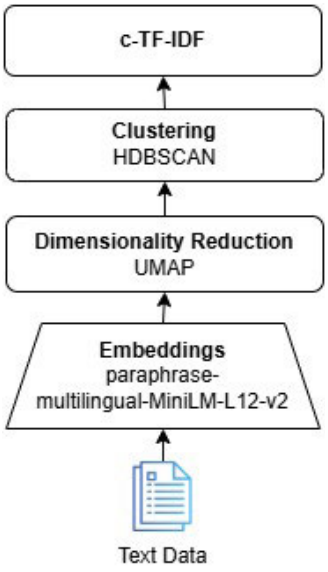


**FIGURE 3.** Audio data processing pipeline with BERTopic.

accuracy and adaptability for complex and nuanced tasks, solidifying its advantage over other technologies.

- **Image Captioning**: The "vit-gpt2-image-captioning" model is used to obtain textual representations of the images, translating the visual details into descriptive language, thereby bridging the gap between visual features and textual analysis. The ViT-GPT2 model combines ViT architecture [42] and a pre-trained GPT-2 language model [49] for generating image captions; recent implementations are given by Hugging Face [50].

#### 3) BERTOPIC(AUDIO DATA)

For the audio component shown in Figure 3, the following process is adopted:

Concurrently, for the audio component $A_i$ of each video $X_i$,

- **Embedding with Multilingual Model** [51]: We use a multilingual embedding model to transform audio transcripts into a rich feature set. By utilizing **paraphrase-multilingual-MiniLM-L12-v2**, a multilingual SBERT [52] variant, we embed the data effectively while preserving the linguistic nuances and context inherent in the transcript. Unlike BERT [53] and RoBERTa [54], these models generate high-quality sentence embeddings directly, with no complex pooling mechanism needed [52]. Meanwhile, BERT [53] is monolingual, and while mBERT [55], though supportive of 104 languages, is not optimized for sentence-level tasks; hence, these are performing so poorly and leading to such high latency. MiniLM-L12-v2, on the other hand, supports over 50 languages [52] and can capture nuances in different linguistic variations in multilingual contexts and hence can handle multilingual transcripts. These reasons make it an ideal choice for our framework, since it gives a good balance between accuracy and multilingual efficiency.

### B. TOPIC EXPLAINABILITY

Topic representations were refined using c-TF-IDF to improve the accuracy and alignment of candidate topics for content taxonomy mapping. To further enhance topic explainability, we conducted evaluations LLMs to assess the coherence and interpretability of the generated topics ensuring stronger alignment with the taxonomy mapping.

We provide prompts in the following customized form:

```
prompt = """
I have a topic that is described by the
   following keywords: [KEYWORDS]
In this topic, the following documents
   are a small but representative subset
    of all documents in the topic:
[DOCUMENTS]

Based on the information about the topic
    above, create a short description of
    this topic with few words.
"""
```

where, particular parameters of the template for the prompt are "[KEYWORDS]" and "[DOCUMENTS]". "[KEYWORDS]" is to be replaced with specific terms representing the topic, whereas in place of "[DOCUMENTS]", a selection from all the documents representative of the topic is to be used. With respect to the visual modality, "[DOCUMENTS]" refers to the image captions generated from the images within each topic.

### C. VIDEO FEATURE EXTRACTION

Video features were extracted by processing audio components, which were obtained from audio transcripts, and visual components, which were obtained from keyframes. Both features were transmitted to textual and visual models for topic inference.

**For visual data**, we extracted keyframes using FFmpeg [56] at a rate of one frame per second (fps = 1) with a similarity threshold exceeding 0.45, ensuring the uniqueness of the captured frames. This method captures the essential visual dynamics of video content effectively. From the visual content of $X_i$, we extracted key images, $V_i$.

**For audio data**, OpenAI's Whisper model [57] was used to obtain accurate multilingual transcripts, with MP4 files prepared using FFmpeg for format compatibility. From the textual content of $X_i$, we extracted the audio transcript, $A_i$.

### D. VIDEO TOPIC INFERENCE

The goal of this step is to pipe through each $X_i$ with both visual and textual BERTopic modeling to get the set of topics identified as $\{T_1, T_2, \ldots, T_n\}$. Each $X_i$ consists of multimodal components: $V_i$ (visual data—key frames) and $A_i$ (audio data—multilingual transcripts). The inference process predicts the topic distribution for each $X_i$, assigning a set of probable topics based on previously learned topic representations.

The predicted topic distribution for the $i$-th video sample can be expressed as:

$$\text{Topic Distribution for } X_i = \{(T_k, p_{i,k}) \mid k = 1, 2, \ldots, n\} \tag{2}$$

where:

- $T_k$: The $k$-th topic, represented by a high-level descriptive label generated by a LLM, summarizing the primary theme or concept of the topic.
- $p_{i,k}$: The probability of the $k$-th topic for sample $X_i$, indicating the relevance of $T_k$ in describing $X_i$. Higher probabilities suggest a stronger alignment between $X_i$ and topic $T_k$.

This probability-based representation enables an interpretable assignment of topics to new data samples, driven from both the visual and textual information. This approach enables the broad understanding of high-level descriptive labels for each topic, which characterizes major themes present in the video sample without referring to individual terms within each topic.

### E. TAXONOMY MAPPING

In the final step of our pipeline, explainable topics were mapped to their closest semantic representations based on relevant items in our content taxonomies. Each entry in the taxonomy [37] includes structured hierarchical information across Tier 1, Tier 2, Tier 3, and Tier 4, listed in separate columns for each row. For the purposes of semantic mapping, we first concatenated these tiers for each row into one hierarchical keyword string, thus capturing the full semantic context of the taxonomy.

For every row $i$, the concatenated taxonomy string $C_i$ is obtained by concatenating nonempty tier values separated by a space:

$$C_i = \text{trim}\left(\sum_{j=1}^{n} \left(T_{i,j} + \text{“”}\right) \cdot \mathbb{I}(T_{i,j} \neq \text{“”})\right) \tag{3}$$

where:

- $T_{i,j}$: The taxonomy term at row $i$ and tier $j$.
- $n$: The total number of tiers (e.g., $n = 4$ for Tier 1 to Tier 4).
- $\mathbb{I}(T_{i,j} \neq \text{“”})$: An indicator function equal to 1 if $T_{i,j}$ is non-empty, and 0 otherwise.
- $+\text{“”}$: Represents the addition of a single space after each tier $T_{i,j}$.
- trim: A function that removes any trailing whitespace in the final concatenated string.

This formula, the tiers are concatenated in sequence from Tier 1 through Tier $n$, in the order of their indices. Further, all the empty tiers are excluded, $T_{i,j} =$“” in the course of the concatenation. Then, a space is added after each non-empty tier to separate the terms in the course of concatenation, and the function 'trim()' is applied to remove those dispensable blank spaces at the end of the concatenated result. The output $C_i$ is therefore a whitespace-free, concatenated string of non-empty tiers from row $i$. This method ensures that $C_i$ is well-formed, free of redundant spaces, and accurately represents the concatenated terms for the given row.

Each of them combines the tiers and uses the resulting strings, $C_i$, in a semantic similarity check against explainable topics. This will, in turn, enable the mapping of each topic to the closest content taxonomy entry according to semantic proximity, hence enabling appropriate topic categorization.

This is computed in terms of cosine similarity, which measures the cosine of the angle between two vectors. The cosine similarity between a topic vector $T_k$ and a taxonomy vector $C_i$ is defined by:

$$\begin{aligned}&\text{cosine\_similarity}(T_k, C_i)\\ &\qquad = \frac{T_k \cdot C_i}{\|T_k\|\|C_i\|},\\ &\qquad \text{where cosine\_similarity}(T_k, C_i) \geq \theta\end{aligned} \tag{4}$$

where:

- $T_k$: The $k$-th topic vector.
- $C_i$: The $i$-th taxonomy vector.
- $\theta$: A predefined cosine similarity threshold, which filters out mappings with scores below $\theta$.
- $T_k \cdot C_i$: The dot product of the vectors $T_k$ and $C_i$.
- $\|T_k\|$ and $\|C_i\|$: The magnitudes of the vectors $T_k$ and $C_i$, respectively.

Finally, the predicted outcome of taxonomies is:

$$\{\mu_1 T_1, \mu_2 T_2, \ldots, \mu_p T_p\} \tag{5}$$

**TABLE 1. Characteristics of the multimodal video analysis dataset.**

| Attribute | Value |
|---|---|
| Number of Videos | 135 |
| Number of Languages | 25 |
| Total Video Duration | 9 hours, 30 minutes, 6 seconds |
| Average Duration per Video | 4 minutes, 10 seconds |
| Number of Taxonomies | 18 |
| Number of Annotated Labels | 188 |

This outcome is obtained based on given term frequencies for each taxonomy. Each taxonomy is associated with a multiplier $\mu_i$ representing term frequency. The multiplier $\mu_i$ indicates the frequency and relevance of the taxonomy based on the mapped topics. In this process, we also used the topic probabilities $p_{i,k}$ derived from Equation (2) during the process of inference and had a threshold to retain only higher-probability topics to map onto taxonomies. For the experiments, it was set to 0.7, where 1 is the highest confidence. The taxonomy frequencies are then counted from these filtered topics, making sure only high-confidence mappings contribute to the final results. This is a parameter that can be set flexibly within the framework to allow a trade-off between quality and coverage in the output. These probabilities represent the taxonomy weights within the framework.

## IV. EXPERIMENTS

### A. TEST DATA

Two distinct datasets were used for training and evaluation. The training dataset, a very rich text corpus, was used to train and fine-tune the textual topic model. A corresponding set of images was used to train and fine-tune the image topic model. The second dataset comprised real-world videos that were used to evaluate the proposed methodology for video taxonomy analysis. Subsequently, fine-tuned models with a rich text corpus and image dataset were applied to a real video dataset to identify and analyze different topics. Model training was focused on the food domain using specialized datasets. For fine-tuning using text, a dataset comprising 180,000 recipes from Food.com obtained from Kaggle was used for a pretrained BERTopic model. In the case of images, the Food101 dataset, consisting of 50,000 images, was used to train a multimodal BERTopic model. The model was tested on YouTube-8M [58], a dataset comprising videos uploaded by users and labeled using the ground truth data. One hundred and thirty-five videos were sampled in the food category as training data, as described in Table 1. These videos contained varying levels of noise, languages, and spacings of time intervals. Wherever necessary, the ground-truth labels were further refined by fixing the accuracy of the updated ground-truth labels. Two experts were enlisted in a user study to assess and select the relevant advertisement taxonomy rows associated with the video content. Each ground-truth label was mapped to a particular row in the taxonomy [37], T1, T2, T3, and T4.

### B. EVALUATION METRICS

Measuring the number of topics in videos using multi-label classification is essential because a video generally involves more than one category. This has motivated considerable research on multi-label classification [59]. Important evaluation measures include Hamming loss, subset accuracy, precision, recall, and F1-score [59], [60].

In multi-label classification, each instance belongs to multiple classes simultaneously. Each binary label can be considered to be a vector $y \in \{0, 1\}^L$, where 1 represents the existence of a specific label from a predefined set $Y = \{\lambda_1, \ldots, \lambda_L\}$, while 0 represents the opposite. For a dataset $D = \{x_1, x_2, \ldots, x_n\}$ consisting of $n$ videos, we consider the task of learning the classifier $h : X \rightarrow Y$ that produces any input in the appropriate sets of labels. In simple terms, $h(x)$ provides a subset of preselected labels for each input, considering an instance with multiple labels. Hamming loss is defined to be the fraction of the number of labels predicted incorrectly. It provides an overall error rate for the classification system [61]. Hamming loss is given by

$$\text{Hamming Loss} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbf{1}_{\{y_{ij} \neq \hat{y}_{ij}\}} \tag{6}$$

where $N$ denotes the number of samples, $L$ denotes the number of labels, $y_{ij}$ denotes the true value of the $j$-th label for the $i$-th sample, $\hat{y}_{ij}$ denotes the predicted value of the $j$-th label for the $i$-th sample, and 1 denotes the indicator function that returns 1 if $y_{ij} \neq \hat{y}_{ij}$ and 0 otherwise.

Subset accuracy is defined to be the number of correctly predicted labels divided by the total number of labels, with a predicted set counted as correct only if it is an exact match of an actual set [62]. It is defined as follows:

$$\text{Subset Accuracy} = \frac{\sum_{i=1}^{N} \mathbf{1}_{\{y_i = \hat{y}_i\}}}{N} \tag{7}$$

where $y_i$ denotes the true label for the $i$-th instance and $\hat{y}_i$ denotes the predicted label for the $i$-th instance.

We use the micro F1-score to optimize the overall matching between content taxonomies in the videos. As the micro F1-score considers the combination of all labels, it is well suited in scenarios with large numbers of labels, and therefore represents classifier performance effectively [63].

Micro-Precision, Micro-Recall, and Micro F1-score:

$$\text{Micro\_precision} = \frac{\sum_{j=1}^{q} \text{tp}_j}{\sum_{j=1}^{q} \text{tp}_j + \sum_{j=1}^{q} \text{fp}_j} \tag{8}$$

$$\text{Micro\_recall} = \frac{\sum_{j=1}^{q} \text{tp}_j}{\sum_{j=1}^{q} \text{tp}_j + \sum_{j=1}^{q} \text{fn}_j} \tag{9}$$

$$\text{Micro\_f1} = \frac{2 \times \text{micro\_precision} \times \text{micro\_recall}}{\text{micro\_precision} + \text{micro\_recall}} \tag{10}$$

where:

- $\text{tp}_j$: True positives for the $j$-th class.
- $\text{fp}_j$: False positives for the $j$-th class.

**TABLE 2.** Optimal parameter values for topic models.

| Topic Model | UMAP Parameters | HDBSCAN Parameters |
|---|---|---|
| **Textual Embeddings** | n_neighbors=100, n_components=5, min_dist=0.0, metric='cosine' | min_cluster_size=200, min_sample=150, metric='euclidean' |
| **Visuals Embeddings** | n_neighbors=3, n_components=4, min_dist=0.0, metric='cosine' | min_cluster_size=30, min_sample=20, metric='euclidean' |

- $\text{fn}_j$: False negatives for the $j$-th class.
- $q$: Number of classes.
- Micro_precision: Micro-averaged precision across all classes.
- Micro_recall: Micro-averaged recall across all classes.

Given by Equations( 8), (9) and (10), these metrics provide an aggregated measure of performance across all classes by treating each instance equally irrespective of its class.

### C. EXPERIMENTAL CONFIGURATIONS: TAXONOMY RETRIEVAL

This involved the tuning of hyperparameters that would enable the training of topic models on the textual and visual content, thus providing the first set of key steps for our experiments. These parameters were iteratively adjusted to capture optimal topic coherence. To ensure a balanced evaluation and avoid overfitting to coherence scores, we fine-tuned BERTopic's hyperparameters to optimize both human-perceived semantic coherence [64] and diversity in topic clusters.

The methodology that was followed consisted of various runs with different hyperparameter settings of UMAP and HDBSCAN for dimension reduction and clustering, respectively, systematically attempted to test the effect they produced on the interpretability and the quality of topics generated. UMAP parameters are: `n_neighbors`, `n_components`, `min_dist`, and `metric`. A list in HDBSCAN also includes `min_cluster_size`, `min_samples`, and `metric`. to tune the cluster for refinement and stability.

The optimal parameters of the text topic model and the visuals topic model, as shown in Table 2, were informed by the nature of the datasets.

This might be explained by the complementary nature of the diverse text corpus and dense image dataset used in this study. Whereas the text corpus provides semantic richness and variation that enables embeddings to capture a wide range of thematic nuances, the dense images ensure that the features are represented in detail, hence fine-grained clustering.

These configurations have balanced interpretability and clustering quality, which meets the model output requirements of the experimental objectives and real-world tasks. Figure 4 summarizes findings for the textual topic model,

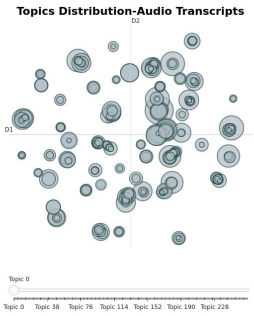


**FIGURE 4.** Topic distribution for audio-based model.

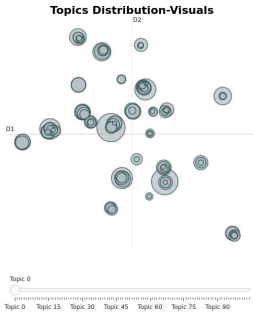


**FIGURE 5.** Topic distribution for visual-based model.

while Figure 5 summarizes the findings of the visual topic model. Both figures illustrate the topic distribution at an optimal point that reflects the coherence achieved with the best configuration for each modality. The found topics were explained with the support of a LLM. In our approach to amplifying LLM selection accuracy, the usage of two different LLMs in our approach was considered:

- **T5-Base model [65]**
- **PaLM 2 model [66]**

Once the model training was complete, the subsequent task was to apply the trained topic models to infer topics from a dataset of 135 selected videos. Additionally, the inferred topics had to be mapped to the predefined content taxonomies such that they would align with the overall thematic structure. We also experimented separately with the unimodal and multimodal methods. These tests enable the comparison of model performance across unimodal and multimodal data and set baselines for further analysis.

- **Unimodal Data**
  - Audio Transcripts data only
  - Visual data only
- **Multimodal Data**
  - Combination of audio transcripts and visual data

IAB content taxonomy [37] are divided into several levels of categorization, namely T1, T2, T3, and T4. Therefore, our experimental design has tried to test the performance of the framework at the higher-order(T1) and finest-grade(T4) levels of the taxonomy spectrum. The results for top-level taxonomies are shown in Table 3, while in Table 4, the results for granular-level taxonomies are given.

**TABLE 3.** Top-level taxonomies inference results using pre-trained topic models configured with PaLM 2 and T5-Base.

| Condition | PaLM 2 model(text-bison-001) | | | | | T5-Base model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-Score |
| Transcript only | 0.54 | 0.13 | 0.57 | 0.55 | 0.56 | 0.53 | 0.13 | 0.55 | 0.53 | 0.54 |
| Visuals Only | 0.43 | 0.15 | 0.59 | 0.80 | 0.68 | 0.43 | 0.15 | 0.59 | 0.80 | 0.68 |
| **Transcript + Visual** | **0.66** | **0.05** | **0.71** | **0.90** | **0.80** | **0.54** | **0.08** | **0.66** | **0.91** | **0.76** |

**TABLE 4.** Granular-level taxonomies inference results using pre-trained topic models configured with PaLM 2 and T5-Base.

| Condition | PaLM 2 model(text-bison-001) | | | | | T5-Base model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-Score |
| Transcript only | 0.21 | 0.17 | 0.55 | 0.31 | 0.38 | 0.21 | 0.15 | 0.52 | 0.32 | 0.36 |
| Visuals Only | 0.22 | 0.17 | 0.62 | 0.53 | 0.57 | 0.22 | 0.17 | 0.62 | 0.53 | 0.57 |
| **Transcript + Visual** | 0.22 | **0.15** | **0.63** | **0.59** | **0.61** | **0.23** | **0.12** | **0.62** | **0.56** | **0.59** |

### D. EXPERIMENTAL CONFIGURATIONS: ROBUSTNESS OF NOISES IN MODALITIES

To ensure that the evaluation of the proposed framework is holistic, one such perspective is constituted by the robustness of the framework under real-world noisy conditions in both visual and transcript modalities. YouTube-8M dataset, which inherently varies to a great degree and is noisy in nature due to being generated by users.

- **Visual Noise**: Irrelevant frames, misaligned content, or videos with very short durations and few representative frames.
- **Noisy Language**: The incomplete transcripts or background music or noise interferes with the speech.
- **Noisy Combination**: Scenarios where the visual and transcript modalities are both noisy at the same time.

In particular, we conducted experiments on a curated subset of 30 videos representing these noise categories from the YouTube-8M dataset. The performance of inferring T1 and T4 taxonomies from videos selected for the noise categories is presented in Tables 5 and 6, respectively.

### E. EXPERIMENTAL CONFIGURATIONS: ROBUSTNESS TO LANGUAGE-SPECIFIC SCENARIOS

Further analysis with respect to the performance of the proposed framework under language-specific conditions is performed by dividing the selected 80 videos into subsets based on the primary language of the transcript. The videos are categorized as:

- **Visuals with English Transcripts**: Videos for which the content of the transcript is completely in English and visually represented.
- **Non-English Transcripts with Visuals**: Videos whose transcript content is mainly composed of non-English languages but come with visual data.

This division enables a focused evaluation of the framework's performance across distinct language-specific scenarios, considering both the availability and quality of transcript data. The results for the T1 and T4 taxonomies, reflecting the effectiveness of multimodal fusion under these conditions, are reported in Tables 7 and 8

## V. RESULT ANALYSIS AND DISCUSSION

This study assesses the efficiency of the proposed framework by quantifying its ability to generate video explainable topics that can be mapped to content taxonomies. These are highlighted in terms of semantic accuracy, strength against noisy conditions, versatility across languages, and qualitative benchmarks against state-of-the-art methods appealing for scalability and industrial-grade implementation.

### A. SEMANTIC ACCURACY IN VIDEO TOPIC MODELING

This reflects that the framework infers the topics of a video in a semantically accurate way and maps them to content taxonomies, as reflected by Tables 3 and 4.

#### 1) TOP-LEVEL TAXONOMIES

From the two sets of evaluation configurations from the topic modeling framework where either *PaLM 2* [66] or *T5-Base* [65] was used for topic labeling the performance is greatly improved by combining transcript and visual inputs. The highest F1-score and subset accuracy are 0.80 and 0.66, respectively, when using the *PaLM 2-configured* model. Whereas the *T5-base-configured* model achieves an F1-score of 0.76 and a subset accuracy of 0.54. The main contribution seen here, particularly of large-parameter models such as PaLM 2, tends to result in semantically richer and more interpretable topic labels that lead to better overall framework performance.

#### 2) GRANULAR TAXONOMIES

At the level of more granular fine-grained distinction and hence more challenging, results of the *PaLM-2-configured topic model* are the F1-score at 0.61, subset accuracy equal to 0.22; while slightly outperforming is *T5-Base-configured* with scores 0.59 and 0.23 correspondingly. At this level, performance decline is reflected by the fact that no detailed topic representation exists in these models trained; hence, further training over more fine-grained data is highly required to have better coverage of topics. Besides, for such a topic modeling task with high accuracy, one can consider hierarchical topic modeling (HTM) [67]. HTM can

**TABLE 5.** Top-level taxonomies inference results under noise conditions using pre-trained topic models configured with PaLM 2 and T5-Base.

* **Visuals (Structured Keyframes)**: Refers to keyframes extracted from selected YouTube-8M videos that are visually clear and informative, providing meaningful visual content. In all other instances, ***Visuals** refers to keyframes extracted from YouTube-8M videos that contain a mix of clear, noisy, or ambiguous visuals, reflecting the natural variability of the dataset. Similarly, ***Transcripts (clean)** refer to well-processed transcripts that accurately capture spoken content and align closely with the video's described content. ***Transcripts** refer to sampled and filtered YouTube-8M video datasets that include a mix of both clear and noisy transcript data.

| Condition | PaLM 2 | | | | | T5-Base | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score |
| Noisy Transcript with Visuals (Structured Keyframes) | 0.51 | 0.11 | 0.70 | 0.81 | **0.77** | 0.47 | 0.13 | 0.65 | 0.78 | **0.71** |
| Noisy Visual with Transcript(clean) | 0.55 | 0.08 | 0.68 | 0.67 | **0.75** | 0.50 | 0.10 | 0.68 | 0.62 | **0.68** |
| Noisy Transcript with noisy Visual | 0.35 | 0.19 | 0.55 | 0.49 | **0.53** | 0.29 | 0.15 | 0.60 | 0.59 | **0.51** |

**TABLE 6.** Granular-level taxonomies inference results under noise conditions using pre-trained topic models configured with PaLM 2 and T5-Base.

| Condition | PaLM 2 | | | | | T5-Base | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score |
| Noisy Transcript with Visual(Structured Keyframes) | 0.23 | 0.19 | 0.60 | 0.52 | **0.57** | 0.21 | 0.26 | 0.45 | 0.38 | **0.46** |
| Noisy Visual with Transcript | 0.19 | 0.19 | 0.56 | 0.48 | **0.46** | 0.13 | 0.24 | 0.47 | 0.42 | **0.41** |
| Noisy Transcript with noisy Visual | 0.01 | 0.29 | 0.25 | 0.22 | **0.28** | 0.01 | 0.29 | 0.22 | 0.21 | **0.23** |

**TABLE 7.** Top-level taxonomies inference results using transcript language with pre-trained topic models configured with PaLM 2 and T5-Base.

| Transcript Condition | PaLM 2 | | | | | T5-Base | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score |
| English Transcripts with Visuals | 0.69 | 0.05 | 0.77 | 0.91 | **0.83** | 0.68 | 0.07 | 0.70 | 0.88 | **0.81** |
| Non-English Transcripts with Visuals | 0.60 | 0.08 | 0.75 | 0.79 | **0.77** | 0.55 | 0.13 | 0.69 | 0.76 | **0.73** |

**TABLE 8.** Granular-level taxonomies inference results using transcript language with pre-trained topic models configured with PaLM 2 and T5-Base.

| Transcript Condition | PaLM 2 | | | | | T5-Base | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score | Subset Accuracy | Hamming Loss | Precision | Recall | F1-score |
| English Transcripts with Visuals | 0.27 | 0.10 | 0.64 | 0.66 | **0.63** | 0.24 | 0.12 | 0.64 | 0.66 | **0.63** |
| Non-English Transcripts with Visuals | 0.25 | 0.12 | 0.59 | 0.63 | **0.56** | 0.22 | 0.15 | 0.55 | 0.52 | **0.50** |

allow finer grainedness and hence better granularity and explainability.

Moreover, these results partly reflect the limitations of using subset accuracy as a measure for multi-label classification. This measure, since it represents only cases where all the taxonomies are predicted correctly, does not account for partially correct predictions, which results in the subset accuracy generally assuming low values [59], [62]. This is particularly reflected in our results at the more difficult granular level, where an exact match is often challenging to get. The F1-score provides a better balance in evaluation for these models.

### B. ROBUSTNESS UNDER NOISY CONDITIONS

The robustness of the framework under different noisy conditions, as shown in Tables 5 and 6, is crucial for real-world applications where data quality is highly variable.

#### 1) TOP-LEVEL TAXONOMIES

On noisy visuals and clean transcripts, *PaLM 2-configured* model achieves a F1-score of 0.75 compared with the baseline score of the *T5-base-configured* model at 0.68. Similarly, for noisy transcripts with clean visuals, the best performance by the *PaLM 2-configured* model secures an F1-score of 0.77 outperforming the *T5-base-configured* model which performed at 0.71. In the most challenging case, both transcripts and visuals are noisy the *PaLM 2-configured* model achieves an F1-score of 0.53, showing that it can maintain reasonable performance even when features are severely limited and noisy. This is slightly better compared to the *T5-base-configured*, which records an F1-score of 0.51 in the same setting.

#### 2) GRANULAR TAXONOMIES

Noise is most evident at the granular level, where fine-grained differentiations are more sensitive to degradation in data. It can be observed that for the noisiest condition, both noisy transcripts and noisy visuals- *PaLM 2-configured* model outperforms at 0.28 while the *T5-base-configured* model at 0.23 F1-score. The performance decline at this level reflects not just the presence of noise but also that this will have a ripple effect due to the unavailability of detailed topic representations in the trained models, as discussed earlier. This limitation is further influenced by the nature of YouTube-8M videos, which are very short in duration, thus providing

limited data for extracting detailed features under the given short time span of the videos. However, since the framework's primary target is identifying ad opportunities, and real-world ads are often positioned after longer content, this issue is likely to be mitigated in real-world conditions, even at a granular level. Despite the overall decrease in accuracy, the framework retains reasonable inference capabilities, highlighting its resilience under the given short time span of the videos.

### C. MULTILINGUAL ADAPTABILITY

As shown in both Tables 7 and 8, the framework handles multilingual input effectively by showing adaptability across non-English transcripts in addition to English.

#### 1) ENGLISH TRANSCRIPTS

The best F1-scores are obtained in the case of English transcripts when supported by visuals, with up to 0.83 at the top level, and 0.66 at the granular level, due to the *PaLM 2-configured* model. These results reflect the inherent optimization that this framework has undergone from primary language, particularly English.

#### 2) NON-ENGLISH TRANSCRIPTS

This demonstrate that it will handle non-English transcripts well, yielding an F1-score of 0.79 at the top level and 0.59 at the granular level when using the *PaLM 2-configured* model. Thus, the system works relatively well on the non-English transcripts, which proves that this approach could be more adaptable with regards to multilingual data with slight degradation in performance.

### D. GENERALIZATION ABILITY OF THE FRAMEWORK

Although the current study focuses on a smaller subset of IAB Tech Lab's content taxonomies, this subset effectively illustrated the capabilities of the proposed approach. Notably, the embedding-based approach of BERTopic applies incremental training to integrate new topics [68] and is therefore naturally scalable for real-world applications where new taxonomies have to be inferred by incremental topic modeling. This reduces the memory needed for training a topic model. For instance, the tested method could scale to cover the full spectrum of IAB Tech Lab's 700+ content taxonomies [69] by processing the data in manageable batches. This scalability is achieved by beginning with initial training on a baseline dataset using the `fit()` method but supports `partial_fit()` updates incrementally [70], thus allowing dynamic learning without full retraining. It remembers previously learned topics while refining or introducing new topics as the data evolves.

Furthermore, for improved explainability of topics and generation of descriptions, the inclusion of models such as PaLM 2 shows reasonable performance gains over T5-Base, while commercially available LLMs like OpenAI's GPT-4 hold even higher promise. While these high-parameter LLMs are expensive to use for inference, our framework restricts their usage to training time, making these inference costs irrelevant for deployment. The approach that makes effective use of such advanced LLMs during training is to use the models saved by these in deployment to come up with accurate and interpretable topic descriptions, improving taxonomy inference without having to use commercial APIs constantly.

Once the model is trained with fine tuned parameters, the inference focuses on two important parameters: topic probabilities $p_{i,k}$ from Equation (2) and cosine threshold $\theta$ from Equation (4). Whereas lower taxonomy weights will allow coverage for niche themes, higher weights will prioritize dominance of topics, hence precision. Similarly, lower cosine thresholds widen the matching with diverse content for wider matching. For instance, general video platform sets lower weights and thresholds for broader coverage, while a specialized provider might want to use higher thresholds for more precision. In our experiments, setting both parameters to 0.7 resulted in an approach that balanced good coverage with reasonable confidence. This underlines the flexibility of the framework to adapt-from general platforms to providers focused on niches-which shows scalability in real-world effectiveness.

### E. COMPUTATIONAL EFFICIENCY DURING INFERENCE

Because the platform can be extended to process the videos in batches, rather than taking the whole video as an input and analyzing it, the computational demands for the inference of long-form video content are minimal. Section V-G discusses this further in detail. This aligns seamlessly with the requirements for ad placements since, on video platforms, ads are placed with a structured frequency to maintain reasonable time blocks. These blocks can be analyzed independently; hence, computational overheads will be reduced and efficiency enhanced. Also, a platform component selection method opted for this purpose provides its full support while maintaining accuracy uncompromised. We adopt the multilingual variant of MiniLM [52] for transcripts, which has a lightweight architecture that ensures computational efficiency [52]. Its latency is significantly reduced compared to BERT and mBERT, while still yielding robust performance for multilingual semantic tasks in terms of bi-text retrieval [52]. For keyframes, we adopt the CLIP-ViT-B/32 model for strong performance balanced with speed. Further, recent improvements, such as Distill-ViT-B/32, have offered improved embedding efficiency using significantly fewer resources [71], although we have not applied this model yet; this might be a promising direction for future work. In the direction of model inference efficiency, serialization of a topic model by the safetensors [72] format minimizes model load times while ensuring safety upon deployment [73]. The experiments were conducted on an `AWS SageMaker notebook` instance of type `ml.t3.xlarge`, which has 4 vCPUs and 16 GB of memory. This setup was good enough

**TABLE 9.** Qualitative comparison of semantic analysis components.

| Aspect | Song et al. [23] | Chong et al. [74] | Singh and Lamba [75] | Proposed Framework |
|---|---|---|---|---|
| **Semantic Understanding** | Employs multimodal fusion with pre-trained models (I3D for action recognition, InceptionV3 for scene analysis, and ResNet for object detection) for action, object, and scene analysis, but remains limited in thematic coherence. | Highlights fusion issues such as obscured modality contributions and suggests modality-specific models like TSN, YOLOv5, CNN14, PlacesCNN. And then leverage Jaccard and Histogram Intersection to align ads to scenes via aggregated relevance scores. | Captions are generated with the help of CNN-LSTM pipelines-for instance, InceptionV3 for visual feature extraction-supplemented with speech recognition for text-based semantic matching and metadata integration but remains limited regarding multimodal robustness | Combines **CLIP-ViT for visuals** and **multilingual embeddings for audio/text**. **Topic modeling** via BERTopic creates coherent semantic clusters that map video content to interpretable, high-level taxonomies. This enhances thematic congruency understanding and aligns ad placements, leading to better encoding and storage, as identified in the literature [8], [76], [77]. |
| **Noise Robustness** | Sensitive to modality-specific noise, studies often emphasize the need for noise-robust multimodal fusion frameworks, as modality-specific noise is a recognized challenge [78], [79]. | The research recognises that noise sensitivity would have been faced upon integrating the place modality into the ad-insertion framework. Authors have discussed how it is expected that the 'place' modality can, by itself, lead to degraded results without complementary modalities on objects. | While CNN-LSTM pipelines exhibit excellent feature extraction capabilities, their performance is highly susceptible to noise. [80]–[82]. | Highly robust: The proposed framework effectively addresses *cross-modality alignment (CMA)* issues, which are common in user-generated video content where audio and video modalities may be semantically misaligned, leading to noisy supervision [83]. While **CLIP-ViT embeddings** make sure that the image and text features align robustly against visual noise [41], the **multilingual embeddings** take care of the noise in the multilingual audio or text content [84]. Additionally, topic modeling filters out the noise through clustering semantically similar embeddings and de-emphasizing outliers in order to attain consistency of results even against the presence of multimodal discrepancies. Thus, it provides robustness for a highly variant noisy dataset. |
| **Explainability** | Abstract feature vectors lack interpretability [85]. | Although it uses semantic features, the relevance scores are derived from mathematical metrics (e.g., Jaccard distance) that are less interpretable than thematic matching [8]. | Traditional methods struggle with explainability in nuanced relationships like synonymy, relying on shallow features [86]. | High: **BERTopic clustering** creates explainable topics with intuitive labels [87]. Mapping topics to taxonomies enhances transparency, enabling stakeholders to understand why specific ads were selected. Topic modeling simplifies interpretation for industry use cases. |
| **Multilingual Capability** | Limited to dataset-specific languages and models. | Language-agnostic features not explicitly supported. | Textual pipelines assume English-centric data, limiting global scalability. | High: **Multilingual embeddings** support audio in more than 25 languages. This topic model is powerful across languages, as it works jointly in an embedding space [88] for consistent performance under multilingual scenarios. |
| **Scalability** | Multiple pre-trained models, along with complex fusion layers and bilinear pooling, increase the computational overhead dramatically, since they need very extensive parameter optimization and processing across modalities [78]. | Extracting high-dimensional features for scenes and advertisements (e.g., action, audio, objects, emotion, and place) using five distinct CNNs can be computationally intensive due to their resource demands. [89]. | High-dimensional vectorization, such as 8000 dimensions, is computationally expensive and inefficient [90] as compared to modern models that are scalable. Both the training and testing of the algorithm are done on the small ADS50 dataset alone, which perhaps is too narrow to represent the diversity of real-world ad pools. | High: Embeddings are lightweight and optimized, will provide reduced computational overhead at runtime. Taxonomy-based topic inference will enable scalability across different types of content, support adaptability to new or evolving taxonomies while the enablement of seamless embedding in advertising ecosystems is highly achievable. |
| **Ad Matching Accuracy** | High when modalities are clean and fused effectively. | High for selected modalities but misses broader thematic contexts. | More ads in a pool increase complexities and ambiguity, and computations potentially reduce accuracy. | High: Assurance about thematic coherence in topic modeling allows ad placements to be contextually aligned. Taxonomy-level fusion makes use of scores about topics to emphasize relevant ad opportunities expected within advertising ecosystems. |

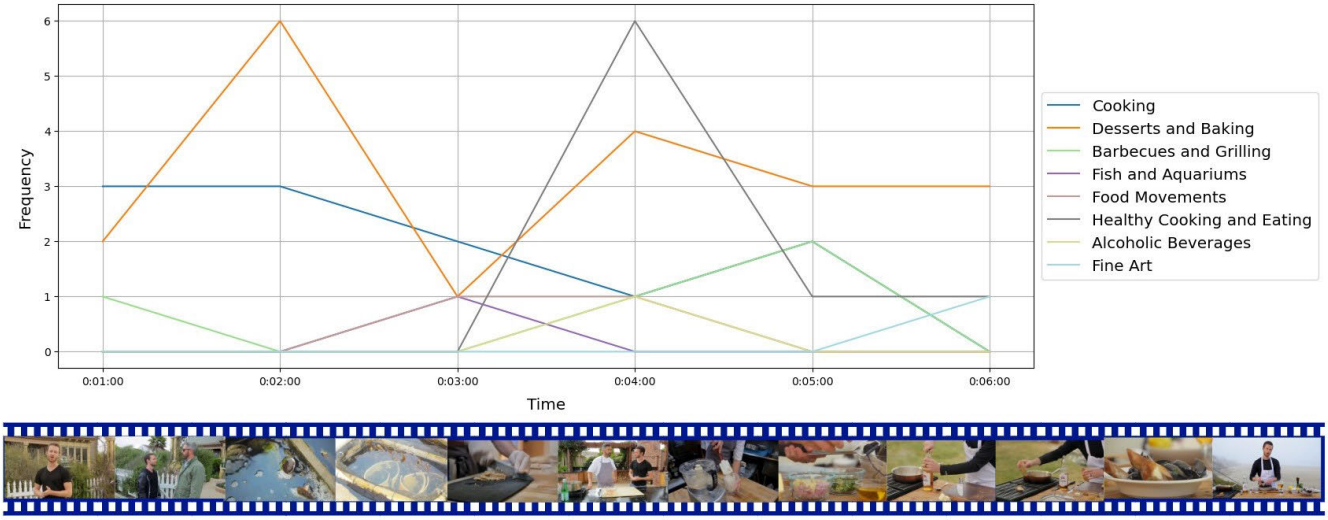


**FIGURE 6.** Evolution of content taxonomy over time in the video.

to perform inference on a 7-minute video for about 3-5 seconds of processing time.

However, Conventional feature extraction from videos, keyframe generation, and transcription are computationally costly tasks. Hence, our optimization in keyframe generation includes the selection of unique frames with a given threshold of similarity 0.45, while our choice is Whisper [57], a lightweight yet accurate multilingual ASR model. These optimizations streamlined the preparation of the input through a reduction in size and complexity of the data and thus allow for efficient and scalable inference at minimal computational costs.

### F. CONCEPTUAL COMPARISON OF SEMANTIC ANALYSIS METHODS FOR CONTEXTUAL ADVERTISING

This section provides a conceptual comparison of the proposed framework against past works in contextual advertising. In order for the comparative analysis to be useful, attention has been restricted to purely semantic-related analysis components of the selected past works, given their direct linkage with the aims of the study. Other factors, such as sentiment analysis or additional auxiliary features considered in prior studies, are outside the scope of this comparison.

The key aspects of semantic understanding, including robustness to noise, explainability, scalability, and multilingual capability, are analyzed conceptually and illustrated in Table 9. These aspects are critical to the proposed framework's ability to deliver explainable, industry-compliant, and globally adaptable contextual advertising solutions. By comparing these elements conceptually, the analysis demonstrates how the proposed approach advances the state-of-the-art methodologies in semantic analysis for contextual advertising.

### G. BROADER FOCUS OF UPCOMING STUDIES

The explainable video topics for content taxonomy framework opens up avenues for a refined contextual advertising solution for long-range videos. By effectively tracking content taxonomies over time, advertisers can dynamically align advertisements with evolving video themes. To address a key challenge in long-form video analysis specifically topic drift where themes evolve or shift over time the framework segments videos into smaller, time-bound units. The initial experiment illustrates that the method of segmentation preserves the coherence of the topic within segments and support the topic drift in longer videos. Furthermore, this strategy aligns with the computational optimizations discussed in V-E, enhancing scalability and efficiency while allowing dynamic adaptation to evolving content.

#### 1) FEASIBILITY SETUP AND INITIAL FINDINGS

Consider a video $X_i$ with the audio transcript $A_i$, consisting of a sequence of phrases $\{p_1, p_2, \ldots, p_m\}$ along with their respective start time codes $\{start_1, start_2, \ldots, start_m\}$ and end time codes $\{end_1, end_2, \ldots, end_m\}$, as well as keyframes $\{k_1, k_2, \ldots, k_n\}$ extracted at specific time intervals within the video. Each keyframe is associated with a timestamp $\{time_1, time_2, \ldots, time_n\}$ that correspond to significant visual content changes in the video.

Video $X_i$ is segmented into several defined time intervals$\{G_1, G_2, \ldots, G_x\}$, where each group $G_i$ represents part of the transcript and the associated keyframes corresponding to the time bounds of that group. For each group $G_i$, the transcript and keyframes $\{p_{i1}, p_{i2}, \ldots, p_{ij}\}$ and $\{k_{i1}, k_{i2}, \ldots, k_{ik}\}$, respectively, were analyzed.

Prediction: The evolution of taxonomies across the segmented groups within the video is represented by

$$\{G_1 : [\mu_{11}T_{11}, \mu_{12}T_{12}, \ldots, \mu_{1p}T_{1p}],$$

$$G_2 : [\mu_{21}T_{21}, \mu_{22}T_{22}, \ldots, \mu_{2p}T_{2p}],$$
$$\ldots,$$
$$G_x : [\mu_{x1}T_{x1}, \mu_{x2}T_{x2}, \ldots, \mu_{xp}T_{xp}]\} \quad (11)$$

where:

- $G_i$ represents each segment of the video.
- $T_{ij}$ is a specific taxonomy identified within group $G_i$.
- $\mu_{ij}$ is the multiplier indicating the frequency and relevance of taxonomy $T_{ij}$ within that specific group.

Thus, the variation of certain taxonomies over a video were mapped to view the thematic progression dynamically. To illustrate the extension of the framework, it is applied to the YouTube "https://www.youtube.com/watch?v=eIcnvKLdxLU" to identify topics evolving in it. The output generated by the taxonomy over time is depicted in Figure 6.

Although these results showed Figure 6 that dynamic taxonomy retrieval was feasible across videos, some areas had more room for improvement. Further video segmentation techniques should be explored in finding logical boundaries without using a fixed time interval, as performed in this feasibility study, for more effective execution. These will incorporate topic transitions and topic-aware sentiment analysis to attain accuracy and relevance for the Taxonomies. Moreover, future research may test these finalized methodologies on long-range video datasets, because a dataset similar to YouTube-8M is incomplete in capturing complexities for a long-range video dataset.

## VI. CONCLUSION

This study proposed a novel framework for the multimodal retrieval approach to industrially compliant contextual advertising through the use of state-of-the-art NLP and multimodal analysis techniques. Transformer-based models, particularly the BERTopic and language models, enable the achievement of video representations aligned with content taxonomies for targeted and relevant advertisements in a privacy-compliant manner. Our methodology will be particularly effective for noisy, multilingual, user-generated content and offer a highly scalable solution for the advertising industry. It will be an increasingly relevant solution as the demand for AVoD services continues to grow, and with traditional behavior-based targeting becoming less effective because of changing privacy regulations. This research contributes to the field of digital marketing by an advancement in techniques of programmatic advertising and hence provides a scalable means of improving ad relevancy, user engagement, and total effectiveness on video platforms.

Building on this research, enhancements in the dynamic tracking of taxonomy changes over time would be a good avenue for future work. Such a development would address one of the most important demands in programmatic advertising: precise ad alignment in the evolution of video content. To this end, refining video segmentation techniques to capture logical content boundaries dynamically is a promising direction. This can be taken further by adding topic-aware sentiment analysis to provide even richer contextual insight into the thematic and emotional subtleties of video content. These efforts go toward making contextual advertising scalable, accurate, and flexible in driving superior user engagement and overall efficiency across video platforms.

## ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers from IEEE Access for their constructive and insightful comments and suggestions which helped in substantially improving the presentation of this article.

## REFERENCES

[1] C. Grece, "Trends in the VOD market in EU28," Eur. Audiovisual Observatory, Strasbourg, France, Tech. Rep., 2021.
[2] S. Broughton Micova and S. Jacques, "Platform power in the video advertising ecosystem," *Internet Policy Rev.*, vol. 9, no. 4, pp. 1–28, 2020.
[3] Z. Liu, U. Iqbal, and N. Saxena, "Opted out, yet tracked: Are regulations enough to protect your privacy?" 2022, *arXiv:2202.00885*.
[4] M. Veale and F. Zuiderveen Borgesius, "Adtech and real-time bidding under European data protection law," *German Law J.*, vol. 23, no. 2, pp. 226–256, Mar. 2022.
[5] T. Mei, J. Guo, X.-S. Hua, and F. Liu, "AdOn: Toward contextual overlay in-video advertising," *Multimedia Syst.*, vol. 16, nos. 4–5, pp. 335–344, Aug. 2010.
[6] T. Mei and X.-S. Hua, "Contextual Internet multimedia advertising," *Proc. IEEE*, vol. 98, no. 8, pp. 1416–1433, Aug. 2010.
[7] T. Kozlova, "Efficiency of business and intercultural communication: Multilingual advertising discourse," in *Proc. III Int. Sci. Congr. Soc. Ambient Intell. (ISC-SAI)*. Amsterdam, The Netherlands: Atlantis Press, 2020, pp. 272–278.
[8] T. Wang and R. L. Bailey, "Processing peripherally placed advertising: The effect of thematic ad-content congruence and arousing content on the effectiveness of in-video overlay advertising," *J. Interact. Advertising*, vol. 23, no. 3, pp. 203–220, Jul. 2023.
[9] S. Segev, W. Wang, and J. Fernandes, "The effects of ad–context congruency on responses to advertising in blogs: Exploring the role of issue involvement," *Int. J. Advertising*, vol. 33, no. 1, pp. 17–36, Jan. 2014.
[10] M. Reisenbichler and T. Reutterer, "Topic modeling in marketing: Recent advances and research opportunities," *J. Bus. Econ.*, vol. 89, no. 3, pp. 327–356, Apr. 2019.
[11] N. Klym and D. Clark, "The future of the ad-supported Internet ecosystem," Tech. Rep., 2019.
[12] J. K. Chalaby, "The streaming industry and the platform economy: An analysis," *Media, Culture Soc.*, vol. 46, no. 3, pp. 552–571, Apr. 2024.
[13] Z. Sherman, *Modern Advertising and the Market for Audience Attention: The U.S. Advertising Industry's Turn-of-the-Twentieth-Century Transition*. Evanston, IL, USA: Routledge, 2019.
[14] A. Turillazzi, M. Taddeo, L. Floridi, and F. Casolari, "The digital services act: An analysis of its ethical, legal, and social implications," *Law, Innov. Technol.*, vol. 15, no. 1, pp. 83–106, Jan. 2023.
[15] T. Mei, L. Yang, X.-S. Hua, H. Wei, and S. Li, "VideoSense: A contextual video advertising system," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 463–464.
[16] K. Okada, E. S. de Moura, M. Cristo, D. Fernandes, M. A. Gonçalves, and K. Berlt, "Advertisement selection for online videos," in *Proc. 18th Brazilian Symp. Multimedia Web*, Oct. 2012, pp. 367–374.
[17] C. Xiang, T. V. Nguyen, and M. Kankanhalli, "SalAd: A multimodal approach for contextual video advertising," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2015, pp. 211–216.
[18] H. Zhang, X. Cao, J. K. L. Ho, and T. W. S. Chow, "Object-level video advertising: An optimization framework," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 520–531, Apr. 2017.
[19] G. Wang, L. Zhuo, J. Li, D. Ren, and J. Zhang, "An efficient method of content-targeted online video advertising," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 40–48, Jan. 2018.
[20] E. Häglund and J. Björklund, "AI-driven contextual advertising: Toward relevant messaging without personal data," *J. Current Issues Res. Advertising*, vol. 45, no. 3, pp. 301–319, Jul. 2024.

[21] R. Tapu, B. Mocanu, and T. Zaharia, "DEEP-AD: A multimodal temporal video segmentation framework for online video advertising," *IEEE Access*, vol. 8, pp. 99582–99597, 2020.

[22] B. Mocanu and R. Tapu, "SemanticAd: A multimodal contextual advertisement framework for online video streaming platforms," *IEEE Access*, vol. 12, pp. 63142–63155, 2024.

[23] X. Song, B. Xu, and Y.-G. Jiang, "Predicting content similarity via multimodal modeling for video-in-video advertising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 569–581, Feb. 2021.

[24] W. Wu, Y. Zhao, Y. Xu, X. Tan, D. He, Z. Zou, J. Ye, Y. Li, M. Yao, Z. Dong, and Y. Shi, "DSANet: Dynamic segment aggregation network for video-level representation learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1903–1911.

[25] S. Jadon and M. Jasim, "Unsupervised video summarization framework using keyframe extraction and video skimming," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 140–145.

[26] V. Truong, "Natural language processing in advertising—A systematic literature review," in *Proc. 5th Asia Conf. Mach. Learn. Comput. (ACMLC)*, Dec. 2022, pp. 89–98.

[27] H. Zhang, Z. Ding, M. Sharid Kayes Dipu, P. Lv, Y. Huang, H. Suleiman Abdullahi, A. Zhang, Z. Song, and Y. Wang, "Identification of illegal outdoor advertisements based on CLIP fine-tuning and OCR technology," *IEEE Access*, vol. 12, pp. 92976–92987, 2024.

[28] Q. Yang, M. Ongpin, S. Nikolenko, A. Huang, and A. Farseev, "Against opacity: Explainable AI and large language models for effective digital advertising," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 9299–9305.

[29] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.

[30] J. Yu, Z. Qin, T. Wan, and X. Zhang, "Feature integration analysis of bag-of-features model for image retrieval," *Neurocomputing*, vol. 120, pp. 355–364, Nov. 2013.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[32] R. Pal, A. A. Sekh, D. P. Dogra, S. Kar, P. P. Roy, and D. K. Prasad, "Topic-based video analysis: A survey," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–34, Jul. 2022.

[33] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.

[34] N. Prakash, H. Wang, N. K. Hoang, M. S. Hee, and R. K.-W. Lee, "PromptMTopic: Unsupervised multimodal topic modeling of memes using large language models," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 621–631.

[35] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the $\beta$-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

[36] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," 2020, *arXiv:2004.07737*.

[37] Inf. Technol. Lab. (2021). *Content Taxonomy 3.0*. Accessed: Jul. 19, 2024. [Online]. Available: https://iabtechlab.com/standards/content-taxonomy/

[38] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[39] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017.

[40] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, vol. 242, no. 1, 2003, pp. 29–48.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[45] S. Cai, L. Qiu, X. Chen, Q. Zhang, and L. Chen, "Semantic-enhanced image clustering," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 6, pp. 6869–6878.

[46] Z. Dang, C. Deng, X. Yang, K. Wei, and H. Huang, "Nearest neighbor matching for deep clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13688–13697.

[47] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. V. Gool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, May 2020, pp. 268–285.

[48] X. Dong, J. Bao, T. Zhang, D. Chen, S. Gu, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "CLIP itself is a strong fine-tuner: Achieving 85.7% and 88.0% Top-1 accuracy with ViT-B and ViT-L on ImageNet," 2022, *arXiv:2212.06138*.

[49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, Feb. 2019.

[50] H. Face. (2023). *Vit-gpt2 Image Captioning*. [Online]. Available: https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

[51] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," 2020, *arXiv:2007.01852*.

[52] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[55] J. Libovický, R. Rosa, and A. Fraser, "How language-neutral is multilingual BERT?" 2019, *arXiv:1911.03310*.

[56] FFmpeg Developers. (2024). *Ffmpeg*. [Online]. Available: https://ffmpeg.org/

[57] OpenAI. (2024). *Whisper*. Accessed: Jul. 19, 2024. [Online]. Available: https://github.com/openai/whisper

[58] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*.

[59] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Inf. Process. Manage.*, vol. 54, no. 3, pp. 359–369, May 2018.

[60] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[61] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State Univ., Corvallis*, vol. 18, no. 1, p. 25, 2010.

[62] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2005, pp. 195–200.

[63] Z. Chase Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," 2014, *arXiv:1402.1892*.

[64] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22, Dec. 2009, pp. 288–296.

[65] H. Won Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[66] R. Anil et al., "PaLM 2 technical report," 2023, *arXiv:2305.10403*.

[67] F. Viegas, A. Pereira, W. Cunha, C. França, C. Andrade, E. Tuler, L. Rocha, and M. A. Gonçalves, "Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics," in *Computational Linguistics*, 2024, pp. 1–59.

[68] N. Gerasimenko, A. Chernyavskiy, M. Nikiforova, A. Ianina, and K. Vorontsov, "Incremental topic modeling for scientific trend topics extraction," in *Proc. Int. Conf. Dialogue*, Jun. 2023, pp. 1–5.

[69] Interact. Advertising Bur. (2024). *Content Taxonomy 3.1*. [Online]. Available: https://iabtechlab.com/standards/content-taxonomy/

[70] T. Agrawal and T. Agrawal, "Solving time and memory constraints," in *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*, 2021, pp. 53–80.

[71] X. Sun, P. Zhang, P. Zhang, H. Shah, K. Saenko, and X. Xia, "DIME-FM: DIstilling multimodal and efficient foundation models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jan. 2023, pp. 15521–15533.

[72] H. F. Safetensors. *Safetensors: A Simple and Safe File Format for Neural Network Weights*. [Online]. Available: https://github.com/huggingface/safetensors
[73] B. Casey, J. C. S. Santos, and M. Mirakhorli, "A large-scale exploit instrumentation study of AI/ML supply chain attacks in hugging face models," 2024, *arXiv:2410.04490*.
[74] O. K. Chong, H.-N. Goh, and J. See, "What modality matters? Exploiting highly relevant features for video advertisement insertion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 3344–3348.
[75] M. Singh and R. Lamba, "Proposing contextually relevant advertisements for online videos," in *Proc. 1st Symp. Mach. Learn. Metaheuristics Algorithms, Appl.*, Trivandrum, India. Cham, Switzerland: Springer, Jan. 2020, pp. 218–224.
[76] D. Davtyan and A. Tashchian, "Thematic congruency in the context of brand placements: Tests of memory and attitude measures," *J. Current Issues Res. Advertising*, vol. 43, no. 3, pp. 319–335, Jul. 2022.
[77] M. Dahlén, S. Rosengren, F. Törn, and N. Öhman, "Could placing ADS wrong be right?: Advertising effects of thematic incongruence," *J. Advertising*, vol. 37, no. 3, pp. 57–67, Sep. 2008.
[78] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.
[79] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion*, Jul. 2020, pp. 1–6.
[80] B. Zhao, C. Cheng, Z. Peng, X. Dong, and G. Meng, "Detecting the early damages in structures with nonlinear output frequency response functions and the CNN-LSTM model," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9557–9567, Dec. 2020.
[81] M. Qiao, S. Yan, X. Tang, and C. Xu, "Deep convolutional and LSTM recurrent neural networks for rolling bearing fault diagnosis under strong noises and variable loads," *IEEE Access*, vol. 8, pp. 66257–66269, 2020.
[82] L. Shang, Z. Zhang, F. Tang, Q. Cao, H. Pan, and Z. Lin, "CNN-LSTM hybrid model to promote signal processing of ultrasonic guided Lamb waves for damage detection in metallic pipelines," *Sensors*, vol. 23, no. 16, p. 7059, Aug. 2023.
[83] J. Wu, Y. Liang, F. Han, H. Akbari, Z. Wang, and C. Yu, "Scaling multimodal pre-training via cross-modality gradient harmonization," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 36161–36173.
[84] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi, and H. Saruwatari, "Learning to speak from text: Zero-shot multilingual text-to-Speech with unsupervised text pretraining," 2023, *arXiv:2301.12596*.
[85] J. Pfau, A. T. Young, J. Wei, M. L. Wei, and M. J. Keiser, "Robust semantic interpretability: Revisiting concept activation vectors," 2021, *arXiv:2104.02768*.
[86] K. A. Nguyen, S. S. I. Walde, and N. T. Vu, "Distinguishing antonyms and synonyms in a pattern-based neural network," 2017, *arXiv:1701.02962*.
[87] D. Kozlowski, C. Pradier, and P. Benz, "Generative AI for automatic topic labelling," 2024, *arXiv:2408.07003*.
[88] E. Zosa and L. Pivovarova, "Multilingual and multimodal topic modelling with pretrained embeddings," 2022, *arXiv:2211.08057*.
[89] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.
[90] A. Rahimi, S. Datta, D. Kleyko, E. P. Frady, B. Olshausen, P. Kanerva, and J. M. Rabaey, "High-dimensional computing as a nanoscalable paradigm," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 9, pp. 2508–2521, Sep. 2017.

**WARUNA DE SILVA** (Graduate Student Member, IEEE) received the B.Sc. and M.B.A. degrees in civil engineering from the University of Moratuwa, Sri Lanka, in 2000 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer and information sciences with the University of Strathclyde, U.K. With over 20 years of experience in the global IT industry, he has excelled in software services, product development, and successful project delivery for enterprise U.S. and U.K. clients. His research interests include natural language processing, machine learning, multimedia data mining, and programmatic advertising.

**ANIL FERNANDO** (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1995, the M.Sc. degree (Hons.) in communications from Asian Institute of Technology, Bangkok, Thailand, in 1997, and the Ph.D. degree in computer science (video coding and communications) from the University of Bristol, U.K., in 2001. He is currently a Professor in video coding and communications with the Department of Computer and Information Sciences, University of Strathclyde, U.K., where he leads the Video Coding and Communication Research Team and also a Visiting Professor with the Center for Vision, Speech and Signal Processing (CVSSP), University of Surrey, U.K. He has been working with all major EU broadcasters, BBC, and major European media companies/SMEs in the last decade to provide innovative media technologies for British and EU citizens. He has graduated more than 110 Ph.D. students and is currently supervising 20 Ph.D. students. He has worked on major national and international multidisciplinary research projects and led most of them. He has published more than 430 papers in international journals and conference proceedings and has published a book on 3D video broadcasting. His main research interests include video coding and communications, machine learning, artificial intelligence, semantic communications, signal processing, networking and communications, interactive systems, resource optimization in 6G, distributed technologies, media broadcasting, and quality of experience.