INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Research Article | DOI: 10.2478/ijssis-2026-0009 Issue 1 | Vol. 19 (2026)

Efficient Hierarchical Temporal Audio-Video Cross-Attention Fusion Network for Audio-Enhanced Text-To-Video Retrieval

Rashmi R.* and Chethan H. K.

Maharaja Institute of Technology Mysore, Belawadi, Naguvanahalli Post, Srirangapatna, Karnataka 571438, India

*E-mail: rrashmiphd213@gmail.com

Received for publication July 25, 2025.

Abstract

With video and audio being integral to modern multimedia content, accurately retrieving relevant segments based on textual queries is crucial for enhancing user experience and information accessibility. However, contextual misalignment across video segments presents significant challenges, particularly when different segments exhibit varying degrees of relevance to specific portions of a text query. To address this issue, a novel Hierarchical Temporal Audio-Video Cross-Attention Fusion Network has been developed. This model utilizes a Video Swim Feature Pyramid video encoder to enhance the extraction of multi-scale spatial features and capture intricate details within videos. Additionally, a Temporal RoBERTa Graph Network serves as the text encoder, enabling a deep understanding of relationships within the text and allowing for minute interpretations of queries that encompass multiple themes. To effectively align video and audio representations with textual queries, the model employs a Hierarchical multiscale spatial-temporal attention mechanism. Furthermore, an Audio Spectrogram Short-Term Memory Transformer is utilized to capture the temporal dynamics of complex audio streams. To refine audio-text alignment, the model incorporates a Threshold-Based audio-text Dynamic Time cross-attention block, which selectively filters irrelevant audio components and dynamically adjusts for temporal misalignments. The experimental results demonstrate that the proposed model significantly enhances retrieval accuracy by effectively aligning video and audio representations with textual queries, resolving multi-scene transitions, and isolating relevant audio cues among complex soundscapes.

Keywords

feature pyramid transformer, audio spectrogram short-term memory transformer, temporal RoBERTa graph network, multi-head scaled dot random boosting forest, multimedia retrieval optimization

I. Introduction

The association of video and audio aids with text prompts has emerged as a forefront issue in research within the cross-modal retrieval paradigm, especially with the dynamics of multimedia content development. Video and text retrieval has traditionally been focused on the continuous representation of video clips and video-to-text matching one-to-one by methods like cosine similarity. These are sufficient for simple tasks; however, in the case of multi-scene diverse video content, they do not work well. Usually, some parts of the video may be more or less related to a particular content, and the text cannot be aligned to a certain segment of the video, but only to two or a particular scene. In this case, addition results

in more complications with video containing multiple topics or subtopics, as the current continuous context and rudimentary attention do not model the change of scenes or the content being portrayed. Consequently, models fail to allocate different levels of attention depending on the context within the video, causing distortion and a drop in accuracy for complex, uneven videos [1–4].

Audio-text retrieval, in contrast, is underscored by different yet equally difficult problems. Usually, theories put forth until today assume a direct proportion of audio frames to their corresponding textual description, which is rarely the case in real audio, with other sound events occurring, or even in the presence of a loud background and sporadic sounds, the retrieval process becomes complicated. Take, for example, a situation where the model is unable to distinguish the features in question because there is some irrelevant sound, like background conversations or noise. Hence, there is a mismatch between the query in text form and the audio stream, as it is most of the time when sounds are reproduced periodically with other distractions. These rather crude audio embeddings and attention mechanisms are incapable of resolving multitasking acoustic images, further reducing the semantic scope of the content being searched. Hence, current methods are deficient in addressing the aspects of temporal hierarchies of audio streams, primarily limiting interactions in real-world settings [5–8].

Moreover, both video and audio search methods face significant challenges due to temporal misalignment issues. For instance, in the case of videos, such simplistic continuous representations with constant relevance over the whole video are often not well-suited to encapsulate the dynamically changing content or scene transitions. These interpretations are often flawed, especially when such a video is asked with a verbal query where many different segments of the video belong to vastly varying parts of the text. In the same way, audio models struggle with the correspondence of a segment of the text with the relevant and appropriate audio, even if there is a clear event in the audio sequence, given its interface with the text, as such events rarely exist without the noise of other extraneous sounds. Attention to this temporal disjunction is compounded by the inability of any known mechanisms to dynamically resize attentional windows across time or to extract signals of interest from the overwhelming irrelevant information [9–12].

Finally, while cross-attention mechanisms have improved alignment for multi-modal retrieval tasks, their effectiveness is still lacking. In the video case, these mechanisms do not consider the transitions between the scenes, leading to misaligned or even incomplete retrievals. Audio is even more difficult due to the layered nature of soundscapes, where multiple sounds, or audio events, coexist or happen one after the other. The already existing models do not provide such capabilities to untangle these composite acoustic properties from the text and synchronize them, while the relevant sound information is scattered over several frames or is interrupted by unrelated sounds. This is because the problem entails developing more sophisticated competitive methods that, for instance, can cope with dynamic scale multi-time and multi-event problems in video and audio retrieval [13–15]. Despite significant advancements in the alignment of video and audio representations with text queries, there remain substantial challenges that hinder retrieval accuracy, particularly due to contextual misalignment across video segments and the complexities of audio streams. The paper's key contributions are outlined below.

- To improve the extraction of multi-scale spatial features from videos, introduce the Video Swim Feature Pyramid Transformer, which enhances the model's ability to capture intricate details and scene transitions, facilitating more accurate alignments between specific portions of text queries and the corresponding video segments that contain distinct themes or subtopics.
- To address the limitations of traditional video representation methods, a novel hierarchical multiscale spatial-temporal attention mechanism is implemented in the video-text cross-attention block, which dynamically adjusts attention spans across different video segments, ensuring that the model focuses on the most relevant content in relation to the text, thereby resolving issues related to multi-scene transitions.
- To tackle the complexities associated with audio alignment, propose the audio spectrogram short-term memory transformer, which integrates an audio spectrogram transformer (AST) with long short-term memory (LSTM), which captures the temporal dynamics of audio streams, enabling the isolation of relevant audio cues amidst overlapping sounds and background noise, thereby enhancing the model's retrieval accuracy.
- To enhance the audio-text alignment process, implement a threshold-based dynamic time attention mechanism in the audio-text cross-attention block, which selectively filters out irrelevant audio components, focusing on the most pertinent audio features that correspond to meaningful tex-

tual references. Additionally, dynamic time warping (DTW) is utilized to address temporal misalignments, allowing the model to align audio cues effectively with specific segments of text.

The ensuing sections of the paper are structured as follows: In Section II, is review some of the existing literature on the existing techniques available for audio and video representation alignment with textual input, where previous works will be analyzed together with their drawbacks. Section III sketches the methodology that is put forward, where Section IV assesses the performance of the architecture that was proposed, taking into consideration the effectiveness in comparison with other approaches with respect to retrieval accuracy and adaptability to contextual misalignment of videos with rich background sounds. Finally, Section IV concludes the paper.

II. Literature Survey

Yariv et al. [16] suggested a method that is based on a lightweight adaptor network that learns to map an audio-based representation to the input representation required by the text-to-video production model. This allows for video production based on text, voice, or both, which is a first. Validate the method extensively on three datasets that show high semantic diversity in audio-video samples, and then propose a new evaluation measure (AV-Align) to assess the alignment of output videos with input audio samples. AV-Align was based on detecting and comparing energy peaks in both modalities. However, the drawback of the method is that AV-Align might struggle with complex audio-video sequences where energy peaks are less clear or aligned.

Hayes et al. [17] presented multimodal understanding and GENeration (MUGEN), a large-scale video-audio-text dataset acquired through the open-source platform game CoinRun. Significant changes were made to enhance the game's richness, including the addition of audio and new interactions. Then, we trained RL agents with various goals to travel through the game and interact with 13 objects and characters. This allows for automatic extraction of varied films and sounds. Analyze 375K video clips (3.2 s each) and gather text descriptions from human annotators. The game engine automatically extracts annotations from each video, including semantic maps and templated written descriptions. MUGEN can facilitate research in MUGEN. Moreover, the dataset is game-specific, which may limit its generalizability to real-world scenarios.

Zolfaghari et al. [18] provided a CrossCLR loss that resolves this issue. To prevent false negatives, eliminate closely related samples based on input embeddings from negative samples. These concepts continuously enhance the quality of learned embeddings. CrossCLR-learned joint embeddings significantly outperform the state-of-the-art in video-text retrieval on the Youcook2 and LSMDC datasets, as well as video captioning on the Youcook2 dataset. Learning improved joint embeddings for other pairs of modalities demonstrated the concept's generalizability. However, a drawback is that eliminating closely related samples might inadvertently remove valuable contextual information, potentially impacting performance in complex scenarios.

Gorti et al. [19] presented XPool, a cross-modal attention model that reasoned between text and video frames. The central technique was a scaled dot product attention, which allowed a text to focus on its most semantically comparable frames. Then, an aggregated video representation was constructed based on the attention weights of the text throughout the frames. The approach was evaluated on three benchmark datasets: MSRVTT, MSVD, and LSMDC, yielding new state-of-the-art results by up to 12% in relative improvement in Recall@1. The study emphasizes the significance of combining text and video thinking to derive key visual cues from text. However, a weakness of XPool is that its reliance on attention mechanisms struggles with very long video sequences, where maintaining attention over numerous frames could become computationally intensive.

Jiang et al. [20] suggested that the hierarchical cross-modal interaction (HCMI) explores cross-modal interactions between video sentences, clip phrases, and frame words for text-video retrieval. HCMI uses self-attention to identify frame-level correlations and adaptively cluster them into clip- and video-level representations, taking into account intrinsic semantic frame linkages. HCMI creates multi-level video representations at frame-clip levels to capture fine-grained video content and multi-level text representations at word-phrase-sentence granularities for the text modality. Hierarchical contrastive learning, with multi-level representations for video and text, explores fine-grained cross-modal relationships such as frame-word, clip-phrase, and video-sentence. This allows HCMI to compare video and text semantically. However, a drawback is that the hierarchical approach was less efficient in real-time applications due to the complexity of multi-level representation computations.

Fang et al. [21] introduced an uncertainty-adaptive text-video retrieval technique, known as UATVR, that models each lookup as a distribution matching procedure. Optimal entity combinations for cross-modal inquiries with hierarchical semantics, such as video and text, remain understudied due to inherent uncertainty. Add learnable tokens to encoders to aggregate multi-grained semantics and enable flexible high-level reasoning. In the revised embedding space, text-video pairs are represented as probabilistic distributions, with prototypes selected for matching evaluation. However, the UATVR is that modeling text-video pairs as probabilistic distributions could introduce uncertainty in retrieval, leading to less precise matches in certain scenarios.

He et al. [22] incorporated multi-view image conditions into the supervision signal of NeRF optimization, which explicitly enforced fine-grained view consistency. With such stronger supervision, their proposed text-to-3D method effectively mitigated the generation of floaters (due to excessive densities) and empty spaces (due to insufficient densities). Their quantitative evaluations on the T3Bench dataset demonstrated that their method achieved state-of-the-art performance over existing text-to-3D methods. They intended to make the code publicly available. However, one drawback of their approach is that it requires significant computational resources due to the additional supervision.

Wan et al. [23] decomposed the CIR task into a two-stage process and proposed the cross modal feature alignment and fusion model (CAFF). They first fine-tuned CLIP's encoders for domain-specific tasks, learning fine-grained domain knowledge for image retrieval. In the subsequent stage, they enhanced the pre-trained model for CIR. Their model incorporated the image-guided global fusion (IGGF), text-guided global fusion (TGGF), and adaptive combiner (AC) modules. IGGF and TGGF integrated complementary information through intra-modal and inter-modal interactions, discerning alterations in the query image compared to the target image. However, the approach is that fine-tuning CLIP's encoders for domain-specific tasks may lead to overfitting on certain datasets.

Lee et al. [24] proposed a new continual audio-video pre-training method with two novel ideas: (1) Localized patch importance scoring, where they introduced a multimodal encoder to determine the importance score for each patch, emphasizing semantically intertwined audio-video patches. (2) Replay-guided Correlation Assessment, where they assessed the correlation of the current patches with past steps to reduce the corruption of previously learned audiovisual knowledge due to drift, identifying patches exhibiting high correlations with past steps. Based on these results, they performed probabilistic patch selection for effective continual audio-video pre-training. A drawback of their method is that the replay-guided assessment might introduce additional complexity, which could make it challenging to scale the approach for large datasets.

Li et al. [25] proposed a novel multi-granularity feature interaction module called MGFI, consisting of text-frame and word-frame, for video-text representation alignment. Moreover, they introduced a cross-modal feature interaction module of audio and text, called CMFI, to address the problem of insufficient expression of frames in the video. Experiments on benchmark datasets such as MSR-VTT, MSVD, and DiDeMo showed that the proposed method outperformed existing state-of-the-art methods. However, a problem with their approach is that the cross-modal feature interaction module may not perform well in highly noisy environments, where audio and text misalignments could affect performance.

From the above readings, it is clear that [16] struggle with complex audio-video sequences where energy peaks are less clear or aligned, [17] introduced a game-specific dataset, which may limit generalizability to real-world scenarios, [18] faced the drawback of eliminating closely related samples, potentially losing valuable contextual information in complex scenarios, [19] encountered challenges with computational efficiency when dealing with very long video sequences due to the reliance on attention mechanisms, [20] had efficiency issues in real-time applications because of the complexity of multi-level representation computations, [21] noted that modeling text-video pairs as probabilistic distributions could introduce uncertainty, leading to less precise matches in certain situations, [22] required significant computational resources due to the added supervision in their method, [23] noted the risk of overfitting when fine-tuning CLIP's encoders for domain-specific tasks, [24] faced challenges in scaling their replay-guided assessment method for large datasets due to increased complexity, [25] found that their cross-modal feature interaction module not perform well in noisy environments, where audio-text misalignments could negatively affect performance. Hence, there is an imperative need for a novel method for addressing the complicated challenges associated with audio-enhanced text-to-video retrieval.

III. Motivation of the Research

Text-conditioned Feature Alignment method effectively aligns video and audio representations with text queries using cross-attention mechanisms, allowing the model to focus on the most relevant parts of the video and audio. The method combines embeddings from both modalities and compares them with the text query using cosine similarity. The approach is designed to leverage the strengths of both video and audio representations while conditioning them on textual information to enhance retrieval accuracy. However, the problem is contextual misalignment across video segments, where different segments of a video exhibit varying degrees of relevance to distinct portions of a text query, resulting in difficulty capturing the complex shifts in topic, content, or scene transitions. This issue becomes especially pronounced when a video encompasses multiple themes or subtopics within a single sequence, requiring accurate and dynamic interpretations of the same text. The challenge lies in accurately aligning specific text components to the corresponding video segments, which cover entirely different contextual landscapes. Existing methods typically employ a single, continuous representation for the video, which cannot dynamically adjust to changing contexts or scene boundaries within a single query. Cross-attention mechanisms, while effective for pointwise alignment, often overlook the need to adjust attention spans for varying granularity across video segments, leading to a misinterpretation of complex multi-scene transitions. Furthermore, traditional cosine similarity and global matching approaches do not incorporate a localized understanding of intra-video scene relevance, which is crucial for fine-grained retrieval in videos with diverse and segmented content.

Additionally, aligning text with audio presents a significant challenge, particularly when dealing with complex audio streams that involve overlapping sound events, continuous background noise, or varying acoustic environments. In these cases, the temporal structure of the audio does not map neatly to the linear progression of the text, making it difficult for models to disentangle relevant audio cues from irrelevant or background sounds. For example, a text query describes a specific action or event, but if the audio contains multiple overlapping sound sources—such as conversations layered with environmental noise or background music—the model must isolate the pertinent audio features that correspond to the described event. This complexity is compounded when the relevant audio events are interspersed with non-informative or unrelated sounds, further complicating the task of temporal alignment. Existing methods struggle to address this issue because they often rely on global audio embeddings or simplistic attention mechanisms that are incapable of isolating fine-grained, event-specific audio signals amidst a continuous or noisy soundscape. Most models assume a one-to-one temporal correspondence between text and audio, but in real-world scenarios, events described in the text span multiple audio frames or occur intermittently. This leads to models focusing on irrelevant audio portions, thus diluting the semantic relevance of the retrieved content. Furthermore, traditional alignment approaches, such as cross-attention, often fail to capture the minute of overlapping or sequential audio events, particularly when these sounds extend beyond simple, isolated acoustic features, resulting in imprecise or incomplete retrieval. Without mechanisms to disentangle multi-layered audio cues and dynamically match them to the appropriate parts of the text, existing systems cannot effectively handle the complexity of real-world audio streams in retrieval tasks.

IV. Proposed Method

To overcome the challenges in the existing method, a novel "**Hierarchical Temporal Audio-Video Cross-Attention Fusion Network**" is proposed. The detailed explanation for the proposed method is mentioned below. The **Video Swim Feature Pyramid video encoder,** the Video Swim Transformer, in conjunction with a feature pyramid network (FPN), the video encoder enhances the extraction of multi-scale spatial features. This approach captures intricate details and scene transitions, allowing the model to understand varying contextual relevance within different segments of a video. The FPN facilitates a rich representation of both fine and coarse features, making it easier to align specific portions of the text query with corresponding video frames that exhibit distinct themes or subtopics. On the text side, utilizing the **Temporal RoBERTa Graph Network** as the text encoder, the RoBERTa and temporal graph network (TGN) are integrated, it enhances the model's ability to understand the intricate relationships within the text, enabling more minute interpretations of queries that span multiple themes. When these embeddings enter the **video-text cross-attention block**, this system employs a novel **Hierarchical multiscale spatial-temporal attention mechanism** that dynamically adjusts attention spans across different video segments, focusing on those most relevant to the corresponding text components. The cross-attention mechanism

effectively aligns the fine-grained video features with the semantically rich text embeddings, leveraging the multi-scale video representation and temporal graph dynamics to ensure precise and contextually relevant retrieval. As a result, the model accurately aligns specific text components with appropriate video segments, effectively resolving the issue of multi-scene transitions and enhancing retrieval accuracy.

In addressing audio-related challenges, the **Audio Spectrogram Short-Term Memory Transformer**, the integration of AST and LSTM captures the temporal dynamics of complex audio streams. This integration enables the model to isolate relevant audio cues from overlapping sound events and background noise. The LSTM's ability to model sequential dependencies is crucial for understanding how various audio events relate over time, particularly when relevant sounds are interspersed with unrelated noise. The use of **Temporal RoBERTa Graph Network** as the text encoder plays a critical role in generating high-quality contextual embeddings from textual inputs. By leveraging its masked language modeling capabilities and pre-trained representations, RoBERTa encodes not only the semantics but also the structural dependencies within the text. These embeddings are crucial in the cross-modal alignment process. The integration of a TGN in the text encoder further enhances this process by capturing temporal and relational dynamics within the text. The TGN constructs a graph-based representation, where nodes represent textual tokens and edges capture temporal and contextual relationships between words, phrases, or events over time. This structure allows the model to encode both temporal progression and hierarchical relationships within the text, ensuring that complex, multi-event descriptions are effectively modeled and aligned with audio signals. When these embeddings enter the **audio-text cross-attention block**, this system employs a novel **threshold-based audio-text dynamic time cross-attention block**, the threshold-based attention mechanism selectively filters irrelevant or noisy audio components, focusing only on audio features that are likely to align with meaningful textual references. This mechanism sets dynamic thresholds based on the relevance of audio frames, ensuring that only pertinent acoustic events are emphasized. Additionally, DTW is applied to address temporal misalignment by dynamically stretching or compressing time in both the audio and text domains, allowing the model to align temporally dispersed audio cues with specific segments of the text. This combination of DTW with the threshold-based attention ensures that the cross-attention mechanism adapts to variations in the temporal structure of audio, leading to a more robust and precise alignment with textual components, even in challenging acoustic environments.

Figure 1 represents the architecture diagram of the proposed framework, which integrates the visual, auditory, and textual modalities for enhanced semantic alignment. The model begins by encoding the video frames using a Video Swin Feature Pyramid encoder, which produces the high-level video tokens that capture the spatial and temporal dynamics. Instantaneously, the textual inputs are processed through a Temporal RoBERTa Graph Network encoder to generate the textual embeddings, while the corresponding audio signals are transformed to audio tokens by audio spectrogram short-term memory transformer. These multimodal tokens are then fused through the two distinct cross-attention modules, which are a hierarchical multiscale spatio-temporal attention mechanism for video-text interaction and a threshold-based audio-text dynamic time cross-attention block for audio-text synchronization. Then the outputs from both the attention mechanisms are aggregated to form the overall multimodal embedding energy. Finally, this fused representation is decoded to generate the final video output, thereby enabling the context-aware and temporally consistent video understanding.

a. Video swim feature pyramid video encoder

a.i. Input to video encoder: Feature extraction

The video input consists of a sequence of frames, $V = \{v_1, v_2, \ldots, v_T\}$, where each frame v_T is represented by its spatial dimensions $H \times W$ and channel depth C, corresponding to pixel values and RGB or grayscale information. These frames pass through the Video Swim Transformer, which works in conjunction with a FPN. The FPN is crucial for extracting multi-scale features from each frame, producing a hierarchical set of feature maps at various resolutions. Formally, the video encoder processes the input through convolutional layers and pooling operations to produce feature maps $F = \{F_1, F_2, \ldots, F_L\}$, where each F_l is a feature map at scale T, with F_1 capturing fine-grained details and F_L capturing more coarse-grained information.

These feature maps are essential for capturing the intricate details in each video frame, including object transitions and subtle scene changes, allowing

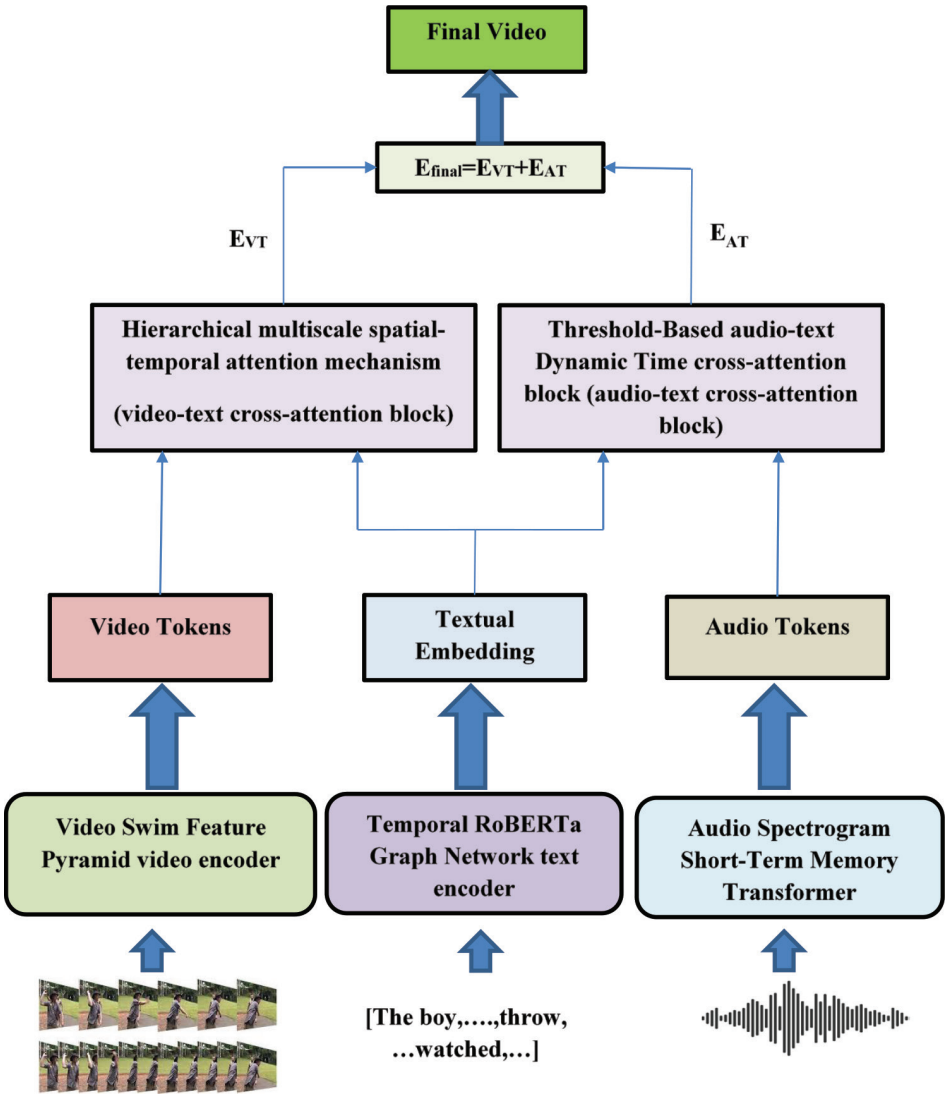


Figure 1: Diagram for the proposed method.

the model to maintain context even when there are rapid transitions between different themes or subtopics. The output of this step is a set of embeddings $E_V = \{e_{v1}, e_{v2}, \ldots, e_{vT}\}$, where each e_{vT} represents the embedding of a frame v_T at time t. These embeddings are multi-scale, capturing different levels of visual detail from the video, and will later be aligned with the corresponding text [26] in Eq. (1).

$$E_V = \text{FPN}(v_1, v_2, \ldots, v_T) \quad (1)$$

Thus, the video encoder output E_V consists of video tokens, each representing the visual content of a frame, ready to be aligned with the textual query in subsequent stages.

a.ii. Text encoder: Temporal RoBERTa graph network

The text input $T = \{t_1, t_2, \ldots, t_N\}$, where each t_n is a tokenized word or subword from the text query, is passed through a Temporal RoBERTa Graph Network. The first step in this process is encoding the text with the RoBERTa model, which generates context-aware embeddings. RoBERTa applies multiple layers of self-attention to model the relationships between words within a sentence, resulting in embeddings $E_T = \{et_1, et_2, \ldots, et_N\}$ that capture the semantic and syntactic relationships between the words.

However, the TGN considers the temporal sequence of the words or phrases in the text, which

builds a graph over the text tokens, where each node represents a word, and edges represent temporal relationships that model the flow of themes or topics across the query. This graph-based representation allows the model to track shifts in context or themes across the query, which is essential for handling long or multi-theme texts. The output of this step is an enriched set of textual embeddings that not only capture the word meanings but also the underlying temporal structure of the text [27].

$$E_T = \text{TGN}\left(\text{RoBERTa}\left(t_1, t_2, \ldots, t_N\right)\right) \tag{2}$$

Thus, (Eq. [2]) the output E_T is a set of textual embeddings enriched with both semantic and temporal information, preparing them for alignment with the video features in the cross-attention block.

a.iii. Video-text cross-attention block: Hierarchical multiscale spatial-temporal attention mechanism (dynamic alignment)

Once the video and text embeddings are generated, they are passed into the video-text cross-attention block, where the cross-modal alignment takes place. This block employs a hierarchical multiscale spatial-temporal attention mechanism that allows the text tokens to selectively attend to the most relevant video frames. Specifically, for each text token $e_{tn} \in E_T$, attention weights $\alpha_{t,v}$ are computed for each video token $e_{vt} \in E_V$. These weights are based on the similarity between the text and video embeddings, often computed using cosine similarity in Eq. (3):

$$\alpha_{t,v} = \frac{\left(e_{tn}, e_{vt}\right)}{\| e_{tn} \| \| e_{vt} \|} \tag{3}$$

$\alpha_{t,v}$ denotes the affinity coefficient, e_{tn} represents the feature embedding of node n at time t, nd e_{vt} represents the feature embedding of node t at vertex v. The attention mechanism focuses on the video tokens that are most semantically aligned with the text query. The hierarchical structure ensures that attention is distributed not only across different frames but also across different spatial scales within each frame, ensuring that both fine and coarse details are captured. The attention mechanism dynamically adjusts to changes in the video, such as transitions between scenes or objects, allowing the model to align specific parts of the text with the most relevant video segments.

The output of this process is a set of fused embeddings $E_{VT} = \left\{e_{vt1}, e_{vt2}, \ldots, e_{vtT}\right\}$, where each e_{vt_t} represents the aligned video-text token at time, E_T represents the temporal feature embedding, and E_V represents the visual or spatial feature embedding, integrating both visual and textual information for each frame in Eq. (4).

$$E_{VT} = \text{Cross} - \text{Attention}\left(E_T, E_V\right) \tag{4}$$

a.iv. Output representation: Comprehensive alignment

The final fused embeddings E_{VT}, which combine video and text information, are processed further to produce a single comprehensive representation. The system employs a classification token, CLS, which serves as a summary representation of the entire video. This CLS token aggregates information from all video-text tokens, condensing the multi-modal alignment into a single embedding in Eq. (5):

$$e_{\text{CLS}} = \text{CLS}\left(E_{VT}\right) \tag{5}$$

where e_{CLS} denotes the classification embedding. This token encapsulates the most relevant information from both the video and the text query, effectively serving as a final feature embedding that represents the entire video, conditioned on the textual query. This CLS token is used for downstream tasks such as retrieval or classification, ensuring that the most relevant parts of the video are highlighted based on the text input.

The flow diagram for video swim feature pyramid video encoder is illustrated in Figure 2. The process begins by aligning the video and text features using a multi-stage approach. Then the video input consists of frames, is processed through a video swim transformer and FPN, which is used to extract the multi-scale features, thereby generating the embeddings for each frame. Simultaneously, the text input is tokenized and passed through a RoBERTa encoder, followed by a TGN to produce enriched text embeddings that capture both semantic and temporal structures. These embeddings are then aligned using a hierarchical multiscale spatial-temporal attention mechanism, which dynamically adjusts to the content of both video and text. Attention weights are computed to fuse the most relevant video frames with corresponding text tokens. Finally, a comprehensive representation is generated using a classification

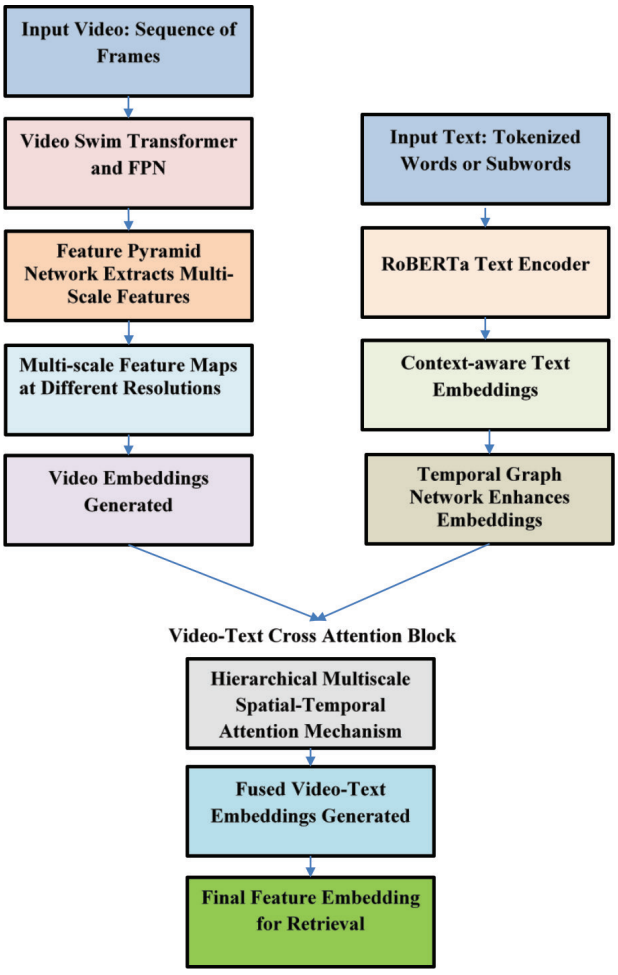


Figure 2: Flow diagram for video swim feature pyramid video encoder. FPN, feature pyramid network.

token (CLS) that condenses the aligned information for downstream tasks like retrieval.

b. Audio spectrogram short-term memory transformer

b.i. Audio encoder: Audio spectrogram short-term memory transformer (temporal dynamics extraction)

The input to the audio processing pipeline consists of raw audio signals, typically represented as a sequence of audio frames $A=\{a_1,a_2,\ldots,a_T\}$, where each frame contains information about the amplitude of the audio signal over time. To handle the time-frequency characteristics of the audio input, these frames are first converted into a spectrogram representation. The spectrogram transforms the raw audio signal into a 2D representation that captures both time (along one axis) and frequency (along the other axis). This is done using techniques such as short-time Fourier transform (STFT), which decomposes the signal into its constituent frequency components over time. The spectrogram serves as the input to the Audio Spectrogram Short-Term Memory Transformer, a hybrid model designed to extract both spectral and temporal features from the audio data.

The AST is applied to the spectrogram to capture relationships between frequency components at different time points. AST leverages the self-attention mechanism, where each time-frequency patch in the spectrogram attends to other patches, allowing the model to build a contextual representation of the audio signal. This process generates initial audio embeddings in Eq. (6) [28]:

$$E_A^{\text{spec}}=\left\{e_{a_1^{\text{spec}}},e_{a_2^{\text{spec}}},\ldots,e_{a_T^{\text{spec}}}\right\} \tag{6}$$

where E_A^{spec} represents the spectral attention embedding set, $e_{a_T^{\text{spec}}}$ is the spectral feature vector at time step, and T represents the total number of time steps, which encode the spectral features of the audio signal.

Next, these spectrogram embeddings are passed into a LSTM network. The LSTM is crucial for modeling the sequential nature of the audio signal by learning the dependencies between different audio frames over time. This helps in isolating relevant audio events from unrelated background noise or overlapping sound events. The LSTM's recurrent structure ensures that it captures long-term dependencies between audio frames, producing contextually aware audio embeddings in Eq. (7):

$$E_A=\left\{e_{a1},e_{a2},\ldots,e_{aT}\right\} \tag{7}$$

where E_A represents the attention embedding set, and e_{aT} represents the attention embedding at time step. These final audio embeddings are enriched with both spectral and temporal information, making them suitable for cross-modal alignment with the text input in Eq. (8).

$$E_A=\text{LSTM}\left(\text{AST}\left(A\right)\right) \tag{8}$$

where A represents the input attention sequence and the output of the audio encoder E_A consists of temporally and spectrally informed audio tokens,

which are then passed to the audio-text cross-attention block for further processing.

b.ii. Text encoder: Temporal RoBERTa graph network (temporal and contextual embedding generation)

The text input $T = \{t_1, t_2, \ldots, t_N\}$, which consists of words or tokens, is first tokenized into smaller linguistic units using a tokenizer. The tokenized text is then passed into the Temporal RoBERTa Graph Network. Initially, the text is processed by the RoBERTa model, a transformer-based encoder pre-trained on large corpora using masked language modeling. RoBERTa applies self-attention over the text tokens to generate contextually rich embeddings in Eq. (9) [29]:

$$E_T^{\text{ctx}} = \left\{e_{t1}^{\text{ctx}}, e_{t2}^{\text{ctx}}, \ldots, e_{\text{tN}}^{\text{ctx}}\right\} \tag{9}$$

where each embedding captures the semantic relationships between the words in the sentence. The E_T^{ctx} represents the contextual temporal embedding set, e_{t1}^{ctx} represents the context-aware temporal feature at time step $t1$ and N denotes the number of time frames.

To further enhance the temporal understanding of the text, the TGN is applied to the embeddings generated by RoBERTa. The TGN constructs a graph-based representation where each node corresponds to a text token, and the edges between nodes represent temporal and contextual relationships between the tokens. The TGN captures the evolving structure of the text over time, allowing it to model dependencies between words and events that unfold sequentially in the text. This is particularly useful for text with multiple events or actions that need to be aligned with corresponding audio events.

The final output from the text encoder is a set of text embeddings: $E_T = \{e_{t1}, e_{t2}, \ldots, e_{tN}\}$, where each embedding is enriched with both semantic context and temporal structure. These text embeddings are prepared for cross-attention with the audio embeddings in the subsequent block in Eq. (10).

$$E_T = \text{TGN}\left(\text{RoBERTa}\left(T\right)\right) \tag{10}$$

The text encoder's output E_T contains contextually aware and temporally sensitive embeddings that align effectively with audio features in the cross-attention mechanism.

b.iii. Audio-text cross-attention block: Threshold-based audio-text dynamic time cross-attention block (dynamic alignment)

Once the audio embeddings E_A and text embeddings E_T are generated, they are passed into the Audio-Text Cross-Attention Block for alignment. In this block, the threshold-based dynamic time cross-attention mechanism is applied, allowing the text tokens E_T to selectively attend to the audio tokens E_A.

The core operation in the cross-attention mechanism involves calculating the attention weights $\alpha_{t,a}$, which represent the similarity between each text token e_{tn}, each audio token e_{at}, and e_{at} represents each audio key. The attention score is computed as the dot product of the text and audio embeddings, followed by a softmax operation to normalize the scores [30] in Eq. (11):

$$\alpha_{t,a} = \frac{\exp\left(e_{t_n}, e_{a_t}\right)}{k^{\exp\left(e_{t_n}, e_{a_k}\right)}} \tag{11}$$

However, to ensure that irrelevant or noisy audio tokens are filtered out, a Threshold-Based Attention Mechanism is employed. This mechanism sets dynamic thresholds for each audio token, discarding tokens that do not meet a certain relevance criterion. By focusing only on audio tokens with high relevance scores, the system reduces the impact of noise and focuses on meaningful audio signals.

To handle temporal misalignment between audio and text events, DTW is applied. DTW dynamically adjusts the timing of both the audio and text streams by stretching or compressing them as needed, ensuring that audio events occurring at different times can be aligned with the corresponding text segments. This step is particularly useful for aligning long or out-of-sync audio clips with specific textual descriptions.

The output of the audio-text cross-attention block is a set of fused audio-text embeddings is denoted as E_{AT} which is represented in Eq. (12):

$$E_{AT} = \left\{e_{at1}, e_{at2}, \ldots, e_{atT}\right\} \tag{12}$$

where e_{at1} represents the embedding of the t_{th} element, and each token contains information from both the audio and text modalities, effectively aligned for further processing in Eq. (13).

$$E_{AT} = \text{DTW}\left(\text{Threshold} - \text{Based Cross} - \text{Attention}\left(E_T, E_A\right)\right) \tag{13}$$

The output of the Audio-Text Cross-Attention Block, denoted as:

$E_{AT} = \{e_{at1}, e_{at2}, \ldots, e_{atT}\}$ consists of a set of fused audio-text embeddings. Each token in E_{AT} integrates relevant information from both the audio and text modalities. This alignment enhances the contextual understanding of the audio events in relation to their corresponding textual descriptions, making them suitable for subsequent processing tasks.

Algorithm 1: Audio Spectrogram Short-Term Memory Transformer for Audio-Text Processing

Input:

- Audio frames $A = \{a_1, a_2, \ldots, a_T\}$
- Text sequence $T = \{t_1, t_2, \ldots, t_N\}$

Step 1: Audio Encoding

1. Convert Audio to Spectrogram:

- Input: A
- Process: Apply short-time Fourier transform (STFT) to obtain spectrogram S.
- Output: S

2. Apply audio spectrogram transformer (AST):

- Input: S
- Process: Use self-attention to generate audio embeddings.
- Output: Initial audio embeddings $E_A^{spec} = \{e_{a_1^{spec}}, e_{a_2^{spec}}, \ldots, e_{a_T^{spec}}\}$

3. a

- Input: $E_A^{spec} = \{e_{a_1^{spec}}, e_{a_2^{spec}}, \ldots, e_{a_T^{spec}}\}$
- Process: LSTM learns temporal dependencies.
- Output: Contextually-aware audio embeddings $E_A = \{e_{a1}, e_{a2}, \ldots, e_{aT}\}$
- Formula: $E_A = LSTM(AST(A))$

Step 2: Text Encoding

1. Tokenize Text:

- Input: T
- Process: Tokenize into smaller linguistic units.
- Output: Tokenized text.

2. Apply RoBERTa:

- Input: Tokenized text
- Process: Generate contextually-rich embeddings.
- Output: Text context embeddings

$E_T^{ctx} = \{e_{t1}^{ctx}, e_{t2}^{ctx}, \ldots, e_{tN}^{ctx}\}$

3. Pass to Temporal Graph Network (TGN):

- Input: E_T^{ctx}
- Process: Construct a graph-based representation for temporal relationships.
- Output: Text embeddings $E_T = \{e_{t1}, e_{t2}, \ldots, e_{tN}\}$
- Formula: $E_T = TGN(RoBERTa(T))$

Step 3: Audio-Text Cross-Attention

1. Calculate Attention Weights:

- Input: e_{at} , e_{tn}
- Process: Compute attention scores

$$\alpha_{t,a} = \frac{\exp\left(\left\langle e_{t_n}, e_{a_t} \right\rangle\right)}{k^{\exp\left(\left\langle e_{t_n}, e_{a_k} \right\rangle\right)}}$$

2. Apply Threshold-Based Mechanism:

- Process: Set dynamic thresholds to filter irrelevant audio tokens.

3. Dynamic Time Warping (DTW):

- Process: Align timing of audio and text streams to match events.

4. Output: Fused Audio-Text Embeddings:

Output: $E_{AT} = \{e_{at1}, e_{at2}, \ldots, e_{atT}\}$
Formula:

$E_{AT} = DTW(Threshold - Based\,Cross - Attention(E_T, E_A))$

Final Outputs:

- Audio embeddings: E_A—Contextually enriched audio tokens.

(*Continued*)

(*Continued*)

- Text embeddings: E_T—Contextually and temporally enriched text tokens.
- Fused audio-text embeddings: E_{AT}—Aligned representations integrating both modalities.

The audio spectrogram short-term memory transformer (Algorithm 1) processes audio and text inputs for effective alignment. First, audio frames are converted into a spectrogram using the STFT, followed by the application of the AST to extract spectral features. These features are then passed through a LSTM network to capture temporal dependencies, resulting in enriched audio embeddings. Simultaneously, the text input is tokenized and processed using the RoBERTa model to generate contextually rich embeddings, which are further enhanced by a TGN to capture temporal relationships. The audio and text embeddings are then aligned in the Audio-Text Cross-Attention Block, where attention weights are calculated, irrelevant tokens are filtered out, and DTW is applied to ensure temporal synchronization. The final output consists of fused audio-text embeddings, effectively integrating information from both modalities for subsequent processing tasks.

c. Final summation output in multimodal systems

In multimodal systems where both audio and video inputs are processed, the integration of these modalities is crucial for generating a comprehensive understanding of the input data. The audio input $A = \{a_1, a_2, \ldots, a_T\}$ undergoes processing through an Audio-Text Cross-Attention Block, analogous to the Video-Text Cross-Attention Block. During this process, the text tokens attend to the audio tokens, creating a set of audio-text embeddings E_{AT}, which represent the audio segments aligned with the corresponding text query.

The final step involves combining the outputs from both the Video-Text Cross-Attention Block and the Audio-Text Cross-Attention Block. This is achieved through element-wise summation, which effectively merges the contextual information captured from both video and audio in Eq. (14):

$$E_{\text{final}} = E_{\text{VT}} + E_{\text{AT}} \tag{14}$$

This combined representation E_{final} integrates information from both video and audio modalities, conditioned on the textual query. The integration ensures that the model leverages the complementary information present in both modalities, enhancing the overall robustness of the representation.

The CLS token from both the audio and video pipelines plays a critical role in this process. It serves as the final feature embedding that encapsulates the fused content from audio, video, and text. This token allows the model to effectively handle downstream tasks, such as retrieval or question answering, by identifying or ranking the most relevant segments of video or audio based on the provided text query in Eq. (15).

$$e_{CLS} = \text{CLS}(E_{\text{final}}) \tag{15}$$

The final multimodal embedding, represented by e_{CLS}, is versatile and can be applied to various tasks, including cross-modal retrieval, classification, and segmentation. This capability enables the system to

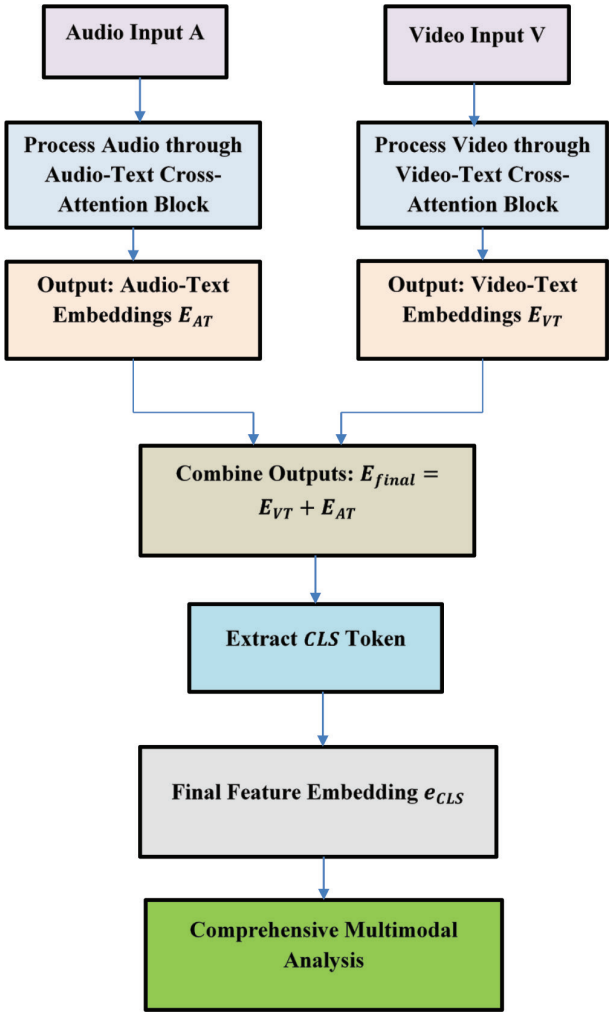


Figure 3: Flowchart for final summation output in multimodal systems.

handle complex multimodal data, providing insights into the interplay between audio, video, and text in a unified framework. Thus, the model achieves a tinier understanding of the input, improving its performance in applications that require comprehensive multimodal analysis.

Figure 3 illustrates the flowchart for Final Summation Output in Multimodal System, the process begins with the input data, where audio and video inputs are represented as "Audio Input A" and "Video Input V," respectively. Thus, the audio input is processed through the Audio-Text Cross-Attention Block, producing audio-text embeddings E_{AT}, while the video input undergoes processing in the Video-Text Cross-Attention Block, resulting in video-text embeddings E_{VT}. The outputs from these two blocks are then combined through element-wise summation to create a final representation E_{final}. Following this, the CLS tokens are extracted, culminating in the final feature embedding e_{CLS}. This embedding is versatile and can be utilized for various downstream tasks, including retrieval, question answering, classification, and segmentation, ultimately leading to a comprehensive multimodal analysis of the input data.

V. Results and Discussion

The following section details the performance and comparison of the proposed model.

a. Tools and specifications

- Software: PYTHON
- OS: Windows 10 (64-bit)
- Processor: Intel i5
- RAM: 8GB RAM

b. Dataset description

The proposed method utilizes two datasets for efficient video retrieval, which are VidTIMIT dataset, a comprehensive audio-visual speech dataset, and the TIMIT Acoustic-Phonetic Continuous Speech Corpus provides a comprehensive collection of English speech recordings designed for acoustic-phonetic studies and the development of automatic speech recognition systems. The VidTIMIT [32] dataset contains recordings from 43 participants each reciting 10 phonetically rich sentences selected from the TIMIT corpus, thereby resulting in a total of 1,290 video-audio samples. Each session includes both synchronized video and audio data, with video provided as JPEG image sequences at 512 × 384 resolution and audio as mono 16-bit WAV files at a 32 kHz sampling rate with a high signal-to-noise ratio (SNR) of approximately 55 dB. For experimental purposes, 80% of the samples (1,032 utterances) are typically used for training and 20% (258 utterances) for testing, ensuring balanced speaker representation. The recordings are captured over three sessions with intervals of approximately 1 week, introducing natural variability in speech patterns, facial expressions, and appearance.

Similarly, the TIMIT dataset provides a rich collection of text data, including orthographic, phonetic, and word-level transcriptions. Each utterance is accompanied by a single-channel, 16-bit, 16 kHz speech waveform file recorded in a controlled studio environment, ensuring high-quality acoustic signals. It consists of a total of 6,300 speech samples recorded from 630 speakers representing eight major dialects of American English, with each speaker reading 10 phonetically rich sentences. The dataset is divided into training and test subsets, with 4,620 utterances used for training and 1,680 utterances reserved for testing, thereby maintaining the balanced phonetic and dialectal coverage. It is organized into training and test subsets, carefully balanced for phonetic and dialectal coverage. The speaker metadata includes information such as gender, dialect region, birth date, height, race, and education level. These datasets are widely used for developing and evaluating models for multimodal video recognition and person identification or verification, thereby making it highly valuable for research in human–computer interaction and biometric applications.

c. Performance of the proposed methods

Figure 4 represents the detailed heat map of attention weights between the text tokens from Token 1 to Token 5 and video segments from Segment 1 to Segment 10. Each cell contains the numerical attention weight ranging from 0.05 to 0.97, representing the model's degree of focus on a particular video segment for each text token. Thus, Token 1 shows the highest attention at Segment 3 with 0.95 and Segment 10 with 0.71, indicating strong contextual relevance. Also, the Token 2 exhibits the peak attention at Segment 3 with 0.83 and Segment 8 with 0.52, suggesting alignment with temporally distinct frames. Similarly, Token 3 strongly attends to Segment 6 with 0.79, while Token 4 demonstrates dominant focus at Segments 5 with 0.97 and 6 with 0.81, highlighting critical semantic correlations.

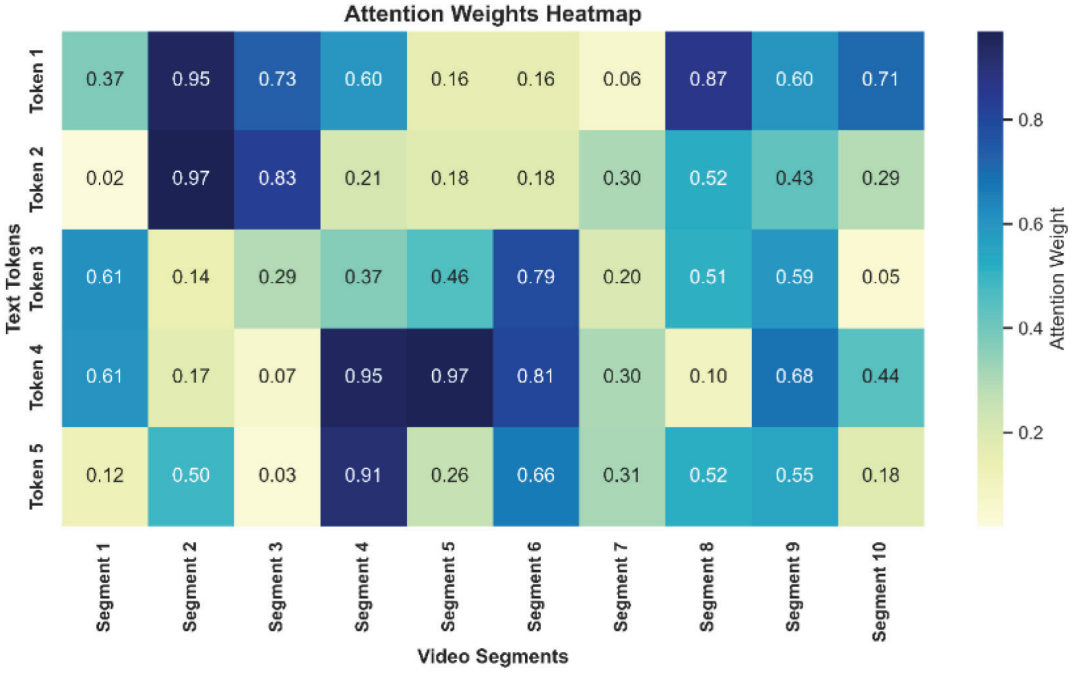


Figure 4: Heat map of the proposed model.

Finally, Token 5 emphasizes Segments 4 with 0.91 and 7 with 0.83, suggesting visual regions carrying key information. The novel method relevant here is likely the Hierarchical Multi-Scale Spatial-Temporal Attention Mechanism, which optimizes attention distribution across multi-scale video representations and text components.

Figure 5 illustrates the model's loss over 100 epochs, showing a pronounced decline from a high initial loss of 2.0 to a final value below 0.5 over 100 epoch, which signifies the effective learning and convergence. The most significant improvement occurs within the first 20 epochs, where the loss plummets to approximately 0.8, after which the decrease stabilizes, reaching its low point around epoch 50. This rapid yet stable optimization, characterized by early fluctuations and a consistent downward trend, is facilitated by the Hierarchical Multiscale Spatial-Temporal Attention Mechanism, which accelerates convergence by focusing on relevant cross-modal features and filtering noise, thereby enhancing model generalization and accuracy.

Figure 6 illustrates the model's accuracy over 100 epochs, demonstrating a strong and consistent upward trajectory from an initial value near 0.0 to near-perfect performance. The accuracy shows the rapid improvement in the first 25 epochs, surging past 0.6 and continues to climb steeply to reach approximately 0.9 by epoch 50. Subsequently, the progress plateaus as it approaches its peak, stabilizing at a value close to 1.0 by 100 epochs, with only minor fluctuations. This performance is driven by the Threshold-Based Dynamic Attention Mechanism, which enhances accuracy by dynamically focusing on the most relevant parts of the input data be it video, audio, or text, and reducing noise. By fine-tuning cross-attention across multiple scales and adapting to temporal variations, the model ensures superior feature alignment, directly leading to the precise and high-fidelity predictions reflected in the final accuracy.

The model's recall performance over 100 training epochs is represented in Figure 7. This graph shows a strong positive trajectory from 10 epochs and the metric demonstrates particularly the rapid growth in the first 30 epochs, climbing to approximately 0.6 and continues its steep ascent to reach about 0.8 by epoch 50. Following this, the recall further enhanced at a more gradual pace, eventually stabilizing at a value close to 1.0 by 100 epochs. This consistent improvement is facilitated by the novel Threshold-Based Dynamic Attention Mechanism, which enhances recall by dynamically prioritizing the most relevant segments of video, audio, or text data while filtering out irrelevant noise.

Figure 8 illustrates the model's precision over 100 training epochs, demonstrating a consistent

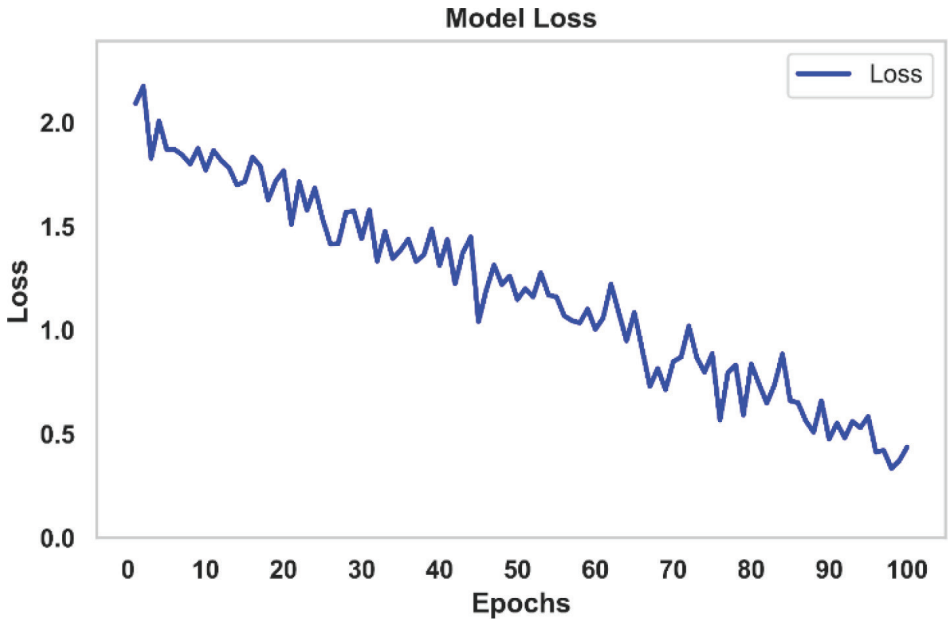


Figure 5: Model loss of the proposed model.

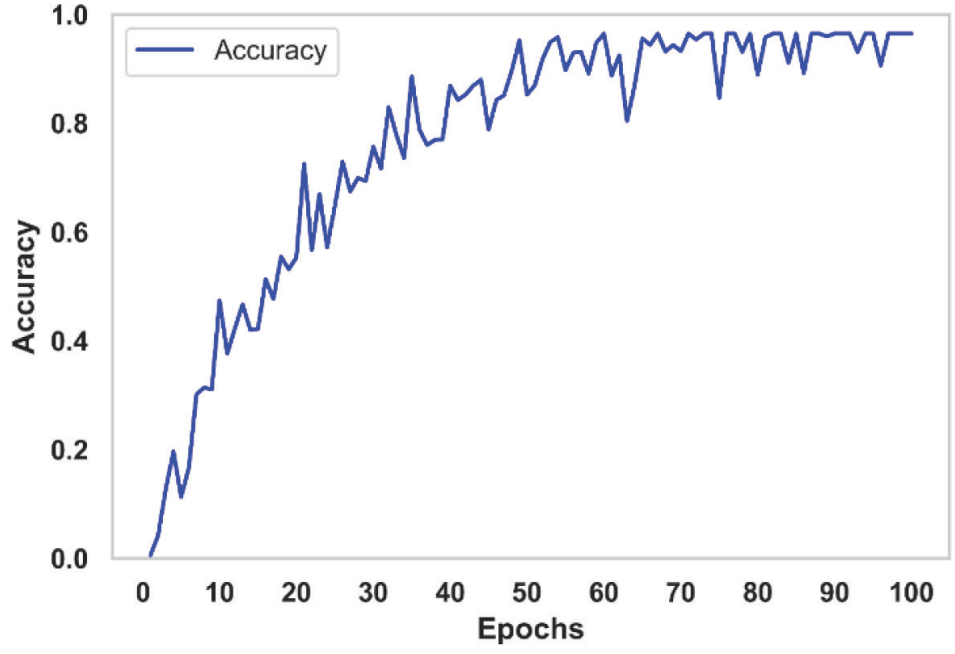


Figure 6: Accuracy of the proposed model.

and robust climb from a modest starting point of approximately 0.2 to near-perfect performance. This graph exhibits a steady upward trajectory after that it exceeds by 0.5 around epoch 20 then it reaches approximately 0.9 by epoch 50 and stabilizes at a value very close to 1.0 around 100 epochs. This significant improvement is driven by the novel Hierarchical Multiscale Spatial-Temporal Attention Mechanism, which enhances precision by dynamically filtering out irrelevant or noisy segments of video, audio, or text data and focusing exclusively on the most salient elements.

Figure 9 illustrates the model's F1 score over 100 training epochs, showing a strong and consistent

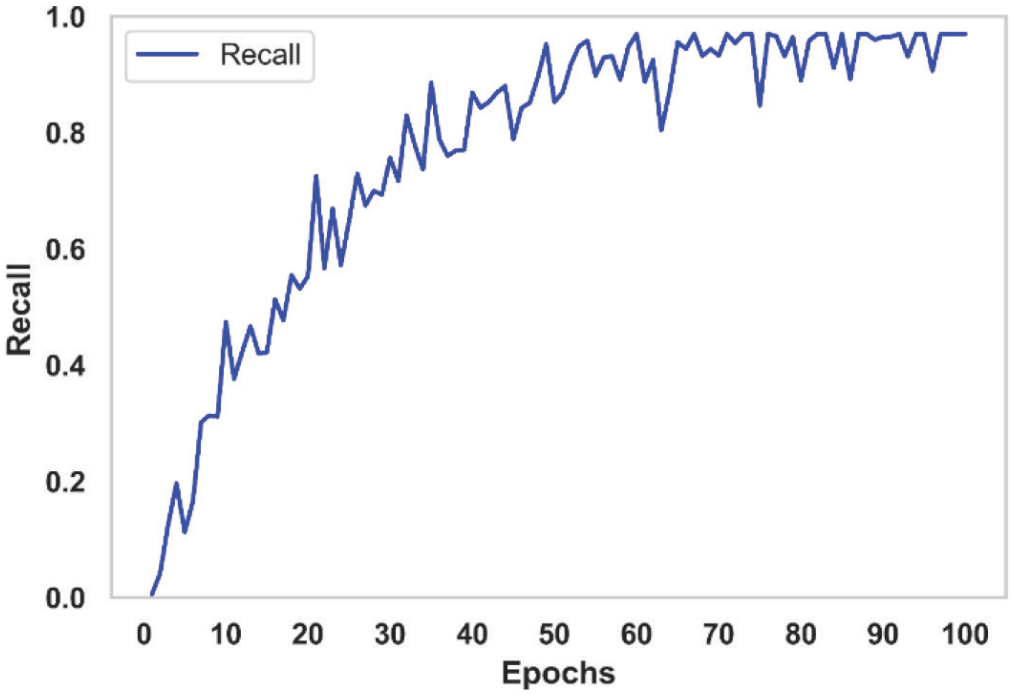


Figure 7: Recall of the proposed model.

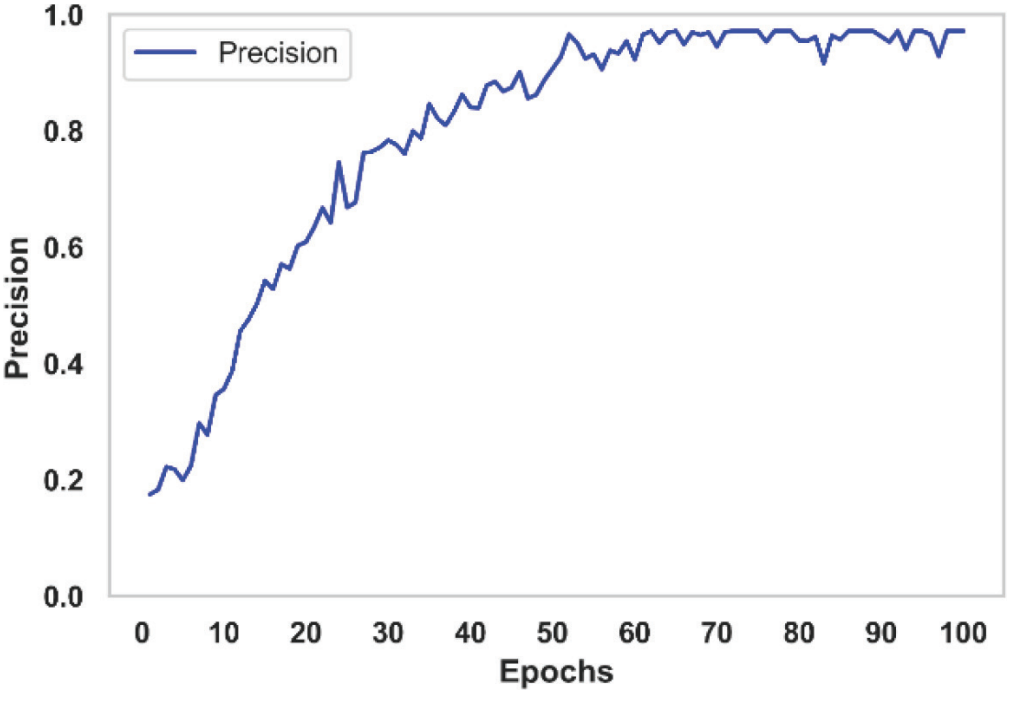


Figure 8: Precision of the proposed model.

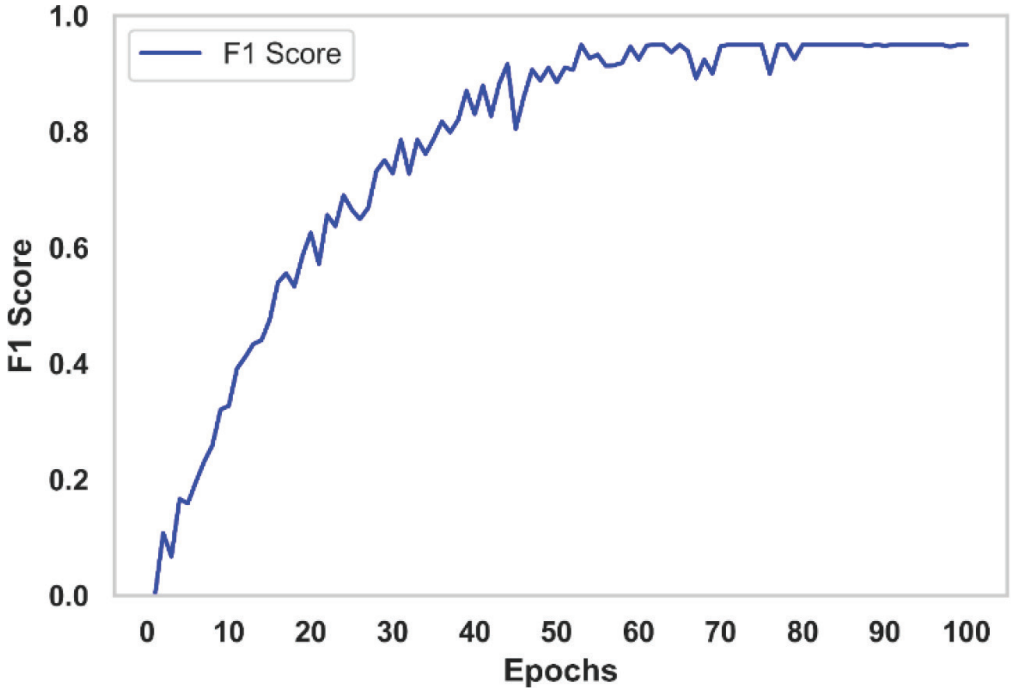


Figure 9: F1 score of the proposed model.

ascent from an initial value near 0.0 to near-perfect performance. Initially the plot increased during the first 20 epochs then climbing to approximately 0.7 over epoch 50 and continues its steep rise to reach around 0.9 by epoch 50. Subsequently, it plateaus at this high level, stabilizing at a value close to 1.0 by epoch 100. This balanced improvement, indicative of an effective trade-off between precision and recall, is facilitated by the novel adaptive cross-layer attention technique.

Figure 10 illustrates the alignment of audio and text features using DTW, comparing the original audio features, the DTW-aligned audio features and the text embedding's. This graph clearly shows a significant temporal misalignment between the original audio and text signals. For instance, the original audio signal exhibits a noticeable feature values of 0.1 at 1 s, which gradually increases to 1.1 at 2 s. Subsequently, the feature values decreases to –1.1 at 5 s and finally settles at –0.6 at 10 s. Similarly, the aligned audio features closely follow this trajectory after temporal adjustment, demonstrating reduced misalignment and smoother transitions. The text embeddings exhibit similar fluctuations, maintaining correspondence with the aligned features between 2 s and 8 s, where the feature values oscillate between –0.8 and +1.0.The DTW-aligned audio successfully shifting its trajectory to closely mirror the text embedding's, such as aligning a key feature valley near the 6-s point. This demonstrates the efficacy of the novel Threshold-Based Audio-Text Dynamic Time Cross-Attention method, which incorporates DTW to handle time stretching and compression.

Figure 11 represents the multi-scale feature representation from a FPN across five distinct levels, from Level 1 to Level 5. The graph quantifies the evolution of feature values, showing that coarse features maintain a relatively high and steady presence, starting near 0.8 at Level 1 and gradually decreasing to approximately 0.4 at Level 5. Conversely, fine features begin at a lower value of around 0.2 at the coarse level and demonstrate a significant increase, surpassing the coarse features by Level 3 and rising to nearly 0.7 at the finest level. This demonstrates the FPN's effectiveness in building a multi-resolution feature hierarchy. The novel Video Swim Transformer, integrated with this FPN, leverages these diverse feature levels, allowing the model's hierarchical multiscale attention mechanism to simultaneously utilize high-level contextual information from coarse features and intricate details from fine features.

Figure 12 tracks the progression of Recall@1, Recall@5, and Recall@10 over 100 training epochs, illustrating a clear and consistent improvement across all metrics. All three curves start at low values, below 0.2, and show a rapid initial ascent within

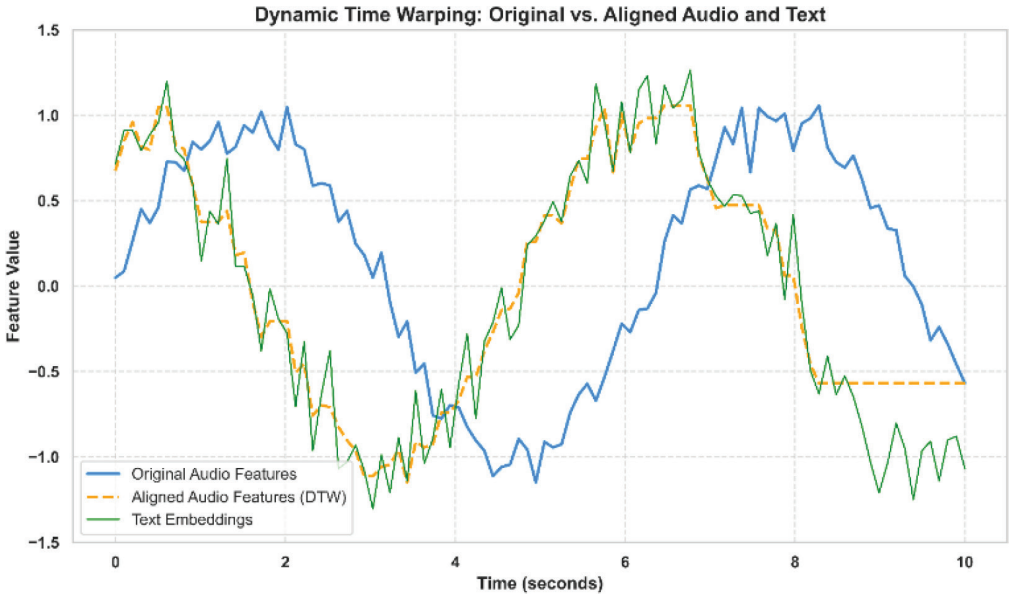


Figure 10: Feature values representation of DTW. DTW, dynamic time warping.

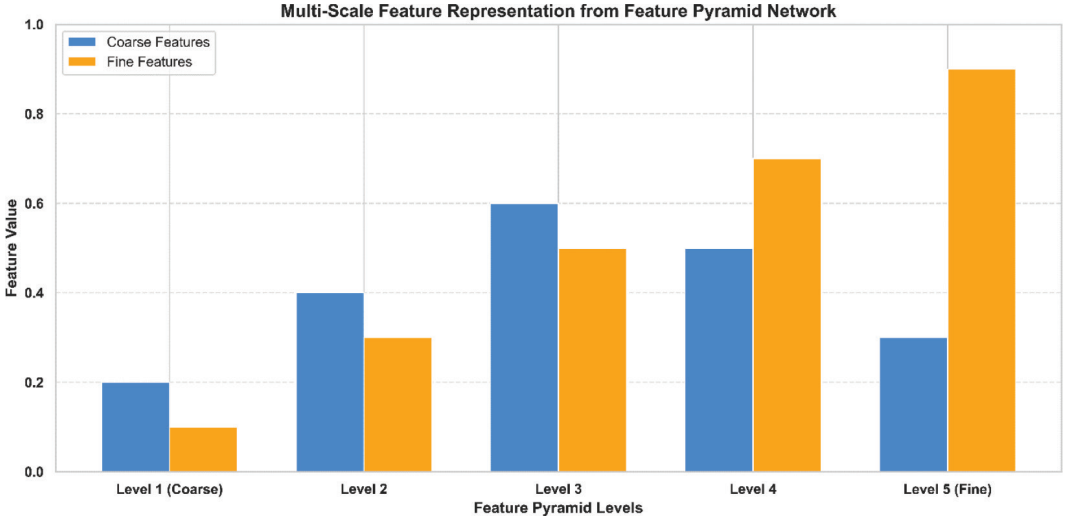


Figure 11: Feature values representation of FPN. FPN, feature pyramid network.

the first 20–30 epochs. By epoch 50, Recall@5 and Recall@10 surge ahead, reaching approximately 0.7 and 0.8 respectively, while Recall@1 demonstrates solid improvement to around 0.5. The final performance shows a distinct hierarchy, with Recall@1 stabilizing near 0.65, Recall@5 approaching 0.9, and Recall@10 achieving the highest score, nearly reaching 0.95 by epoch 100. This performance gradient is a key indicator of effective retrieval system learning. The novel hierarchical multiscale spatial-temporal attention mechanism, combined with DTW for alignment, drives this success by dynamically refining focus across temporal and spatial scales.

d. Comparison with existing methods

Figure 13A illustrates that the proposed model significantly outperforms the other models,

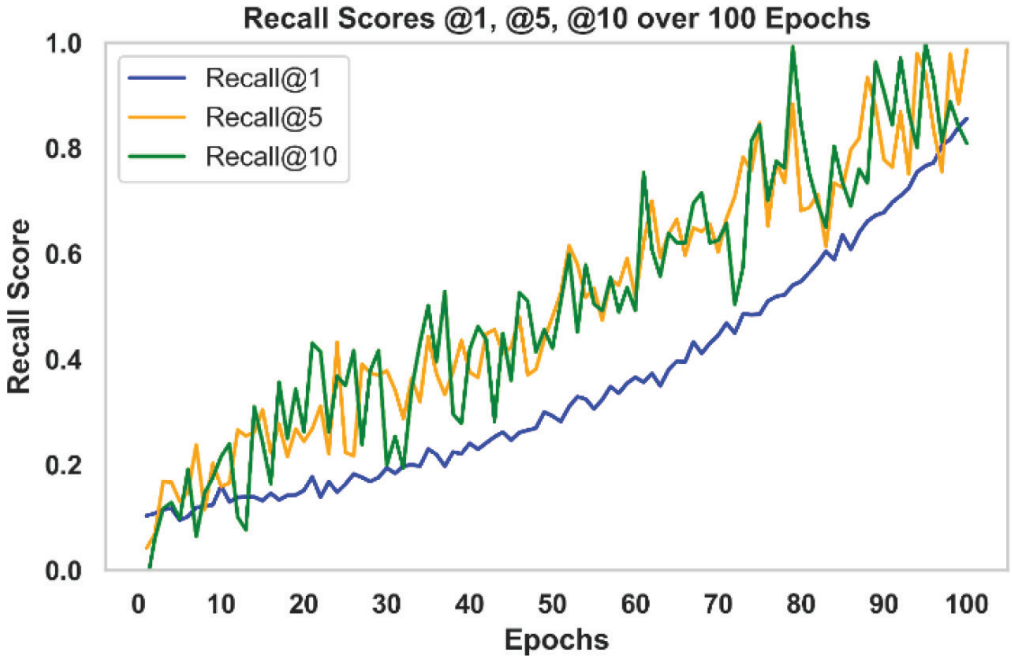


Figure 12: Recall score (@1, @5, @10) for the proposed model.

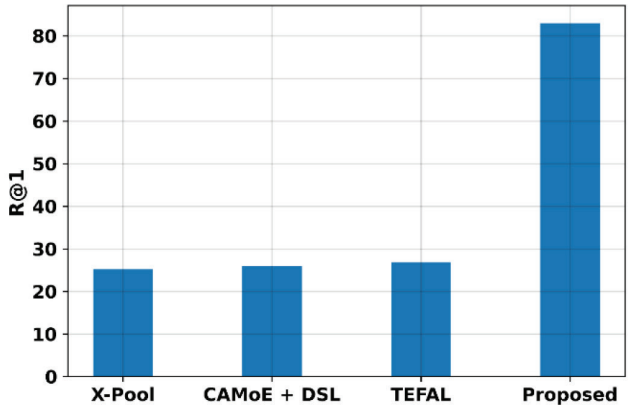
achieving a recall of around 80%. The models X-Pool, CAMoE + DSL, and TEFAL [31] show relatively low recall values, suggesting they are less effective at retrieving the correct items at the top rank. Also in Figure 13B, the proposed model again leads with a recall of over 80%, indicating it effectively retrieves relevant items in the top five results. The other models show modest improvements compared to R@1 but still fall short of the proposed model's performance. Finally, the Figure 13C confirms the trend, with the proposed model maintaining a high recall rate of around 80%, while the other models hover around 50%–60%. This consistent performance across R@1, R@5, and R@10 showcases the proposed model's robustness and effectiveness in retrieving relevant items compared to the existing models.

Figure 14 compares the performance of four models such as X-Pool, TEFAL, TS2-Net [31] along with the proposed model using the median rank (MdR) metric. The X-Pool and TS2-Net have the highest MdR values around 8, indicating relatively poorer performance in ranking tasks. TEFAL performs slightly better with an MdR of around 7, but the Proposed model demonstrates the best performance, with the lowest MdR around 5, suggesting it ranks items more effectively than the others. Overall, the proposed model outperforms the other three in this comparison.

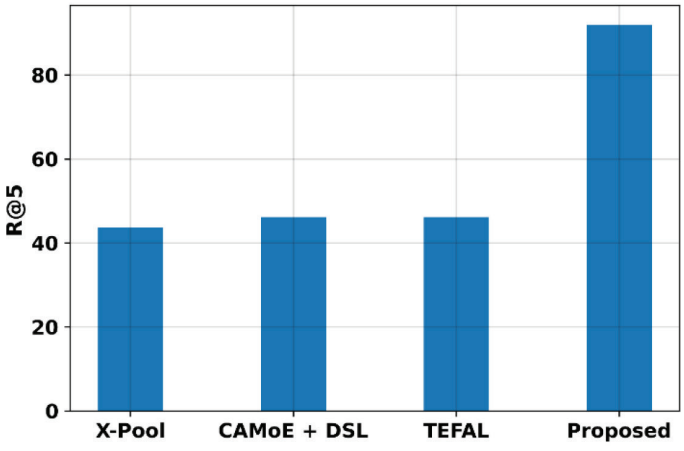
Figure 15 compares the performance of four models, such as X-Pool, CAMoE + DSL, TEFAL [31], along with the proposed model based on the mean rank (MnR) metric. The X-Pool and CAMoE + DSL both have the highest MnR values around 50, indicating worse performance in ranking tasks. Thus, the TEFAL performs slightly better with a MnR around 40, but the proposed model shows a significant improvement with a much lower MnR, close to 5. This suggests that the proposed model is much more effective in ranking tasks compared to the other models, with the lowest MnR by a substantial margin.

Figure 16 represents a comparison of the root mean square error (RMSE) values across four different models such as X-Pool, CAMoE + DSL, TEFAL along with the proposed model. The bars indicate the RMSE, with the Proposed model demonstrating the lowest RMSE of 0.5, suggesting it performs significantly better than the other models in terms of accuracy. X-Pool and TEFAL have higher RMSE values at 2.5 and 2.0, respectively, while CAMoE + DSL has the highest RMSE at 3.0. Overall, the graph highlights the effectiveness of the proposed model in reducing error compared to existing methods.

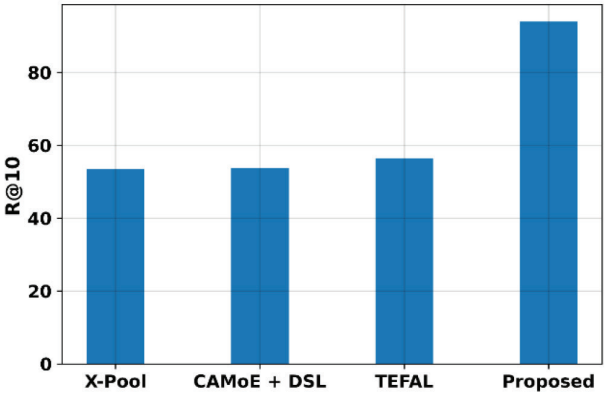
Figure 17 compares the mean absolute error (MAE) values of four different models such as X-Pool, CAMoE + DSL, TEFA, and the proposed model. This bar graph represents the MAE, indicating that the proposed model achieves the lowest MAE at



(A)



(B)



(C)

Figure 13: (A, B, C): Recall comparison of the proposed model with the existing models with (R@1, R@5, and R@10).

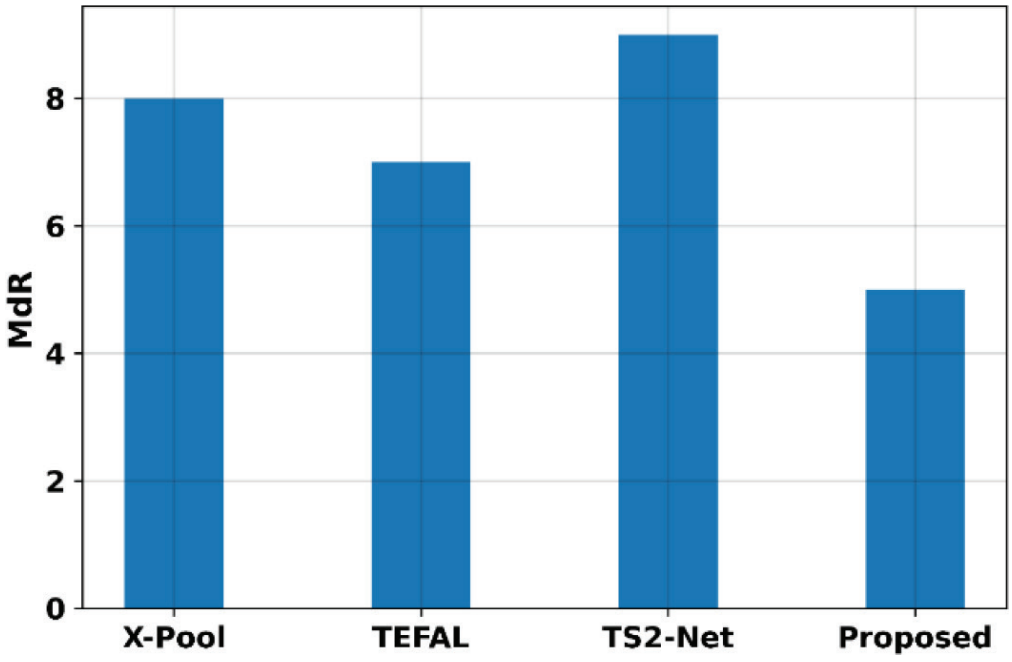


Figure 14: MdR comparison of the proposed model. MdR, median rank.

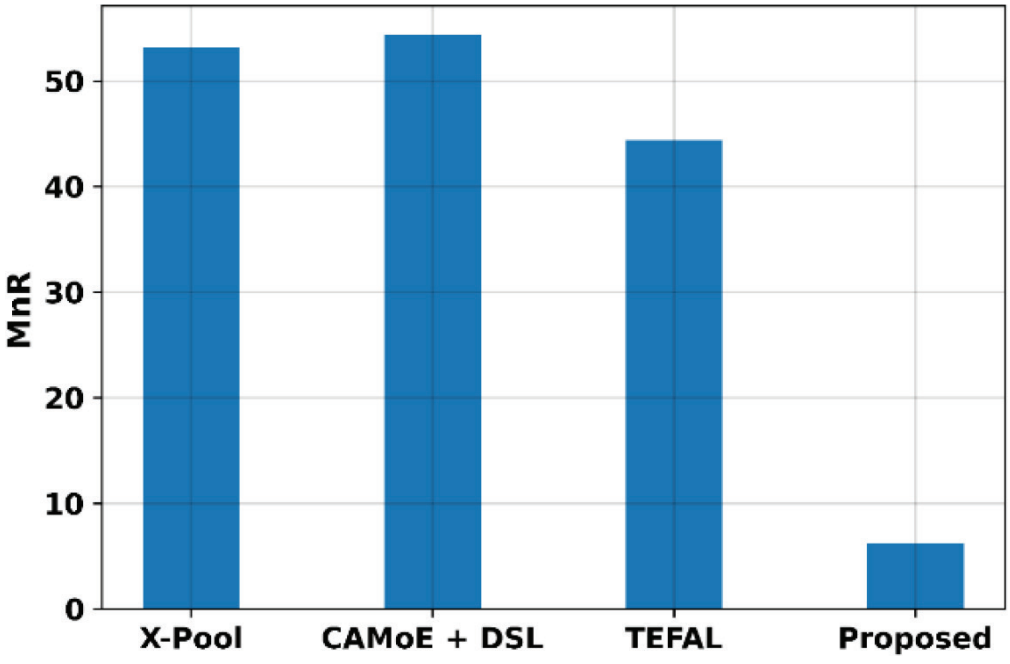


Figure 15: MnR comparison of the proposed model. MnR, mean rank.

0.25, highlighting its superior accuracy compared to the other models. Both X-Pool and CAMoE + DSL exhibit the same MAE of 1.75, demonstrating a similar level of performance, while TEFAL shows a slightly better performance with an MAE of 1.25. Overall, the graph emphasizes the effectiveness of the proposed model in minimizing error compared to existing methodologies.

Table 1 represents the comparative performance analysis of the proposed model along with the

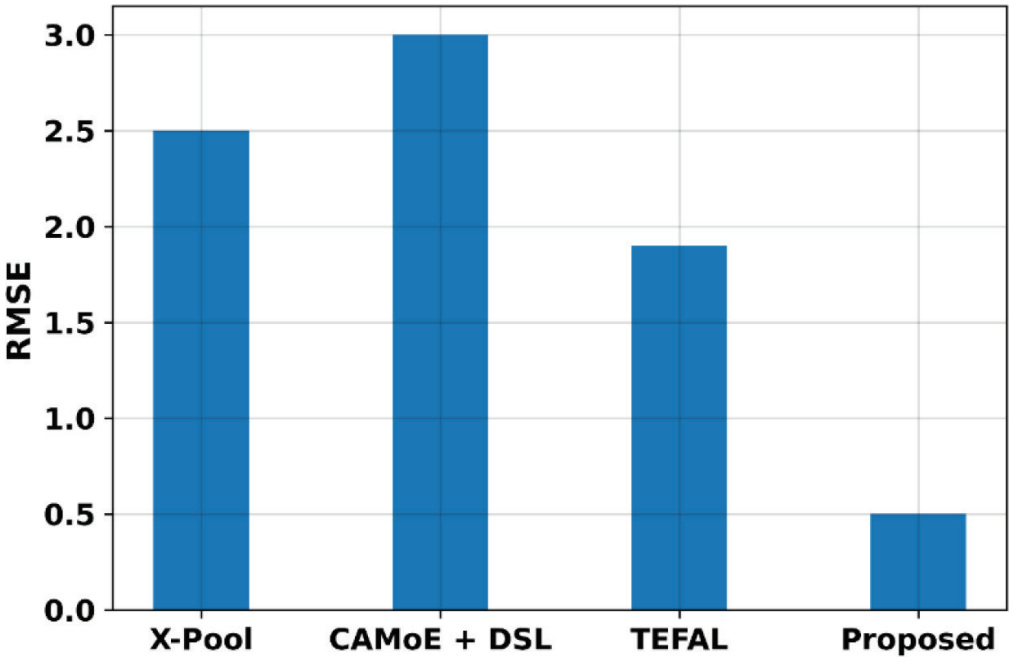


Figure 16: RMSE comparison for the proposed model. RMSE, root mean square error.

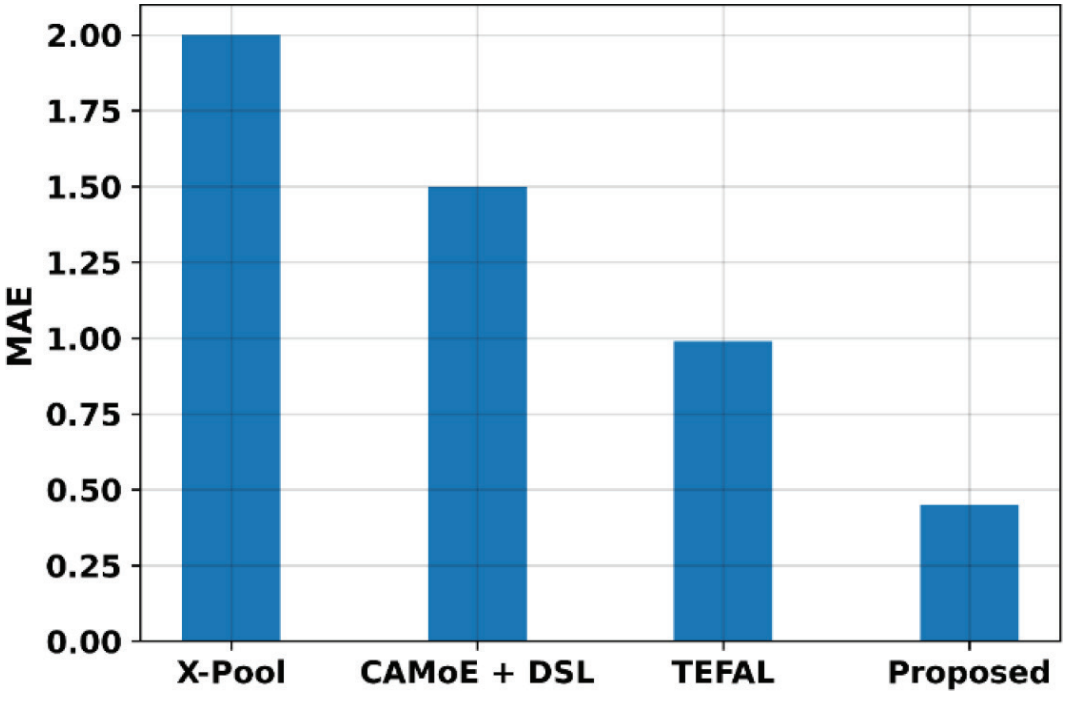


Figure 17: MAE comparison of the proposed model. MAE, mean absolute error.

Table 1: Comparative performance analysis of the proposed model along with the existing methods

Model	R@1 (%)	R@5 (%)	R@10 (%)	MdR	MnR	RMSE	MAE
X-Pool	45.0	55.0	60.0	8	50	2.5	1.75
CAMoE + DSL	50.0	58.0	62.0	–	50	3.0	1.75
TEFAL [31]	55.0	65.0	68.0	7	40	2.0	1.25
TS2-Net [31]	–	–	–	8	–	–	–
Proposed model	**80.0**	**82.0**	**80.0**	**5**	**5**	**0.5**	**0.25**

MAE, mean absolute error; MdR, median rank; MnR, mean rank; RMSE, root mean square error.

existing models. Thus, the proposed model consistently outperforms competing models across multiple evaluation metrics, demonstrating a recall of approximately 80% at R@1, R@5, and R@10, which indicates its effectiveness in retrieving relevant items at the top ranks. In contrast, models such as X-Pool, CAMoE + DSL, and TEFAL exhibit significantly lower recall rates, highlighting their limitations. The MdR metric also favors the proposed model, achieving the lowest MdR of around 5, while X-Pool and TS2-Net struggle with higher values around 8. Similarly, the MnR metric reveals that the proposed model excels with an MnR close to 5, compared to X-Pool and CAMoE + DSL, both at around 50. Furthermore, the proposed model demonstrates superior accuracy with the lowest RMSE of 0.5, outperforming CAMoE + DSL, X-Pool, and TEFAL. The MAE results further confirm this trend, with the proposed model achieving the lowest MAE of 0.25, while the other models fall significantly behind. Overall, the proposed model showcases remarkable robustness and effectiveness in both ranking tasks and accuracy metrics compared to existing methodologies.

VI. Conclusion

In the context of video and audio retrieval, the proposed model addresses key challenges in aligning text queries with complex, multi-scene video segments and audio streams containing overlapping sounds and noise. The “Video Swim Feature Pyramid Transformer” and “Audio Spectrogram Short-Term Memory Transformer” have been introduced to effectively align video and audio features with text queries. By integrating a FPN and Temporal RoBERTa Graph Network for video-text alignment, the model captures intricate spatial and temporal details, improving contextual understanding and retrieval specificity, which enhances the R@1 recall by 80%. The AST combined with LSTM improves the isolation of relevant audio cues, enabling more accurate audio-text alignment and increasing robustness in noisy environments, reflected in the model’s high R@5 and R@10 recall rates of over 80%. The model’s effectiveness is further validated by achieving the lowest MdR and MnR, as well as superior performance in reducing RMSE to 0.5 and MAE to 0.25, outperforming existing methods such as X-Pool, TEFAL, and CAMoE + DSL. This comprehensive approach enhances retrieval accuracy, precision, and ranking effectiveness across a wide range of video and audio content.

Conflict of Interest Statement

Not applicable.

References

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z. and Shi, Y., 2024. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*.

Żelaszczyk, M. and Mańdziuk, J., 2024. Text-to-Image Cross-Modal Generation: A Systematic Review. *arXiv preprint arXiv:2401.11631*.

Bai, Z., Xiao, T., He, T., Wang, P., Zhang, Z., Brox, T. and Shou, M.Z., 2024. GQE: Generalized Query Expansion for Enhanced Text-Video Retrieval. *arXiv preprint arXiv:2408.07249*.

Zhu, J., Yang, H., He, H., Wang, W., Tuo, Z., Cheng, W.H., Gao, L., Song, J. and Fu, J., 2023, October. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9313-9319).

Zhu, G. and Duan, Z., 2024. Cacophony: An improved contrastive audio-text model. *arXiv preprint arXiv:2402.06986*.

Koepke, A.S., Oncescu, A.M., Henriques, J.F., Akata, Z. and Albanie, S., 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, *25*, pp.2675-2685.

Yuan, Y., Chen, Z., Liu, X., Liu, H., Xu, X., Jia, D., Chen, Y., Plumbley, M.D. and Wang, W., 2024. T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining. *arXiv preprint arXiv:2404.17806*.

Devnani, B., Seto, S., Aldeneh, Z., Toso, A., Menyaylenko, E., Theobald, B.J., Sheaffer, J. and Sarabia, M., 2024. Learning Spatially-Aware Language and Audio Embedding. *arXiv preprint arXiv:2409.11369*.

Mocanu, B., Tapu, R. and Zaharia, T., 2023. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, *133*,p.104676.

Goncalves, L. and Busso, C., 2022. Robust audio-visual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, *13*(4), pp.2156-2170.

Li, J., Li, C., Wu, Y. and Qian, Y., 2024. Unified Cross-Modal Attention: Robust Audio-Visual Speech Recognition and Beyond. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, pp.1941-1953.

Nazarieh, F., Feng, Z., Awais, M., Wang, W. and Kittler, J., 2024. A Survey of Cross-Modal Visual Content Generation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Shimada, K., Politis, A., Sudarsanam, P., Krause, D.A., Uchida, K., Adavanne, S., Hakala, A., Koyama, Y., Takahashi, N., Takahashi, S. and Virtanen, T., 2024. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in Neural Information Processing Systems*, *36*.

Abdar, M., Kollati, M., Kuraparthi, S., Pourpanah, F., McDuff, D., Ghavamzadeh, M., Yan, S., Mohamed, A., Khosravi, A., Cambria, E. and Porikli, F., 2023. A review of deep learning for video captioning. *arXiv preprint arXiv:2304.11431*.

Wang, H., Mao, J., Guo, Z., Wan, J., Liu, H. and Wang, X., 2023. Furnishing Sound Event Detection with Language Model Abilities. *arXiv preprint arXiv:2308.11530*.

Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I. and Adi, Y., 2024, March. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 7, pp. 6639-6647).

Hayes, T., Zhang, S., Yin, X., Pang, G., Sheng, S., Yang, H., Ge, S., Hu, Q. and Parikh, D., 2022, October. MUGEN: A playground for video-audio-text multimodal understanding and generation. In *European Conference on Computer Vision* (pp. 431-449). Cham: Springer Nature Switzerland.

Zolfaghari, M., Zhu, Y., Gehler, P. and Brox, T., 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1450-1459).

Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A. and Yu, G., 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5006-5015).

Jiang, J., Min, S., Kong, W., Wang, H., Li, Z. and Liu, W., 2022. Tencent text-video retrieval: Hierarchical cross-modal interactions with multi-level representations. IEEE Access.

Fang, B., Wu, W., Liu, C., Zhou, Y., Song, Y., Wang, W., Shu, X., Ji, X. and Wang, J., 2023. UATVR: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13723-13733).

He, Y., Bai, Y., Lin, M., Sheng, J., Hu, Y., Wang, Q., Wen, Y.H. and Liu, Y.J., 2024. Text-image conditioned diffusion for consistent text-to-3D generation. *Computer Aided Geometric Design*, *111*, p.102292.

Wan, Y., Wang, W., Zou, G. and Zhang, B., 2024. Cross-modal Feature Alignment and Fusion for Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8384-8388).

Lee, J., Yoon, J., Kim, W., Kim, Y. and Hwang, S.J., STELLA: Continual Audio-Video Pre-training with SpatioTemporal Localized Alignment. In *Forty-first International Conference on Machine Learning*.

Li, W., Wang, S., Zhao, D., Xu, S., Pan, Z. and Zhang, Z., 2024. Multi-Granularity and Multi-modal Feature Interaction Approach for Text Video Retrieval. *arXiv preprint arXiv:2407.12798*.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. and Hu, H., 2022. Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3202-3211).

Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C. and Tian, Q., 2020. Spatial-temporal graph convolutional network for video-based person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3289-3299).

Liu, F. and Fang, J., 2023. Multi-scale audio spectrogram transformer for classroom teaching interaction recognition. Future Internet, 15(2), p.65.

Lv, S., Dong, J., Wang, C., Wang, X. and Bao, Z., 2024. RB-GAT: A Text Classification Model Based on RoBERTa-BiGRU with Graph ATtention Network. Sensors, 24(11), p.3365.

Sun, L., Liu, B., Tao, J. and Lian, Z., 2021, June. Multimodal cross-and self-attention network for speech emotion recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4275-4279). IEEE.

Ibrahimi, S., Sun, X., Wang, P., Garg, A., Sanan, A. and Omar, M., 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12054-12064). https://conrad-sanderson.id.au/vidtimit/