Computer Engineering and Applications Vol. 14, No. 3, October 2025

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

Rashmi R., Chethan H.K.

Maharaja Institute of Technology Mysorea
rrashmiphd213@gmail.com, hkchethanphd213@gmail.com

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

1. 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 two

video, whole videos, or in particular scenes and segments. In this case, addition results in more complications with video containing multiple topics or subtopics as 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 make an assumption of 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 whereby 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 the times 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 primary limiting interactions in the 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 synch 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.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

- To improve the extraction of multi-scale spatial features from videos, introduce the Video Swim Feature Pyramid Transformer, 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 Audio Spectrogram Transformer with 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, selectively filters out irrelevant audio components, focusing on the most pertinent audio features that correspond to meaningful textual 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 2, 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 4 sketches the methodology that is put forward, whereon Section 5 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 6 concludes the paper.

2. FORMATING INSTRUCTION

Author Guy et al [16] suggested method 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.

Thomas et al [17] presented MUGEN (Multimodal Understanding and GENeration), 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 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 seconds 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 multimodal understanding and generation. Moreover, the dataset is game-specific, which may limit its generalizability to real-world scenarios.

Mohammadreza 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.

Satya 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 was struggle with very long video sequences, where maintaining attention over numerous frames could become computationally intensive.

Jie et al [20] suggested 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.

Bo 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.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

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.

Yuze 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 completely 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 require significant computational resources due to the additional supervision.

Yongquan et al [23] decomposed the CIR task into a two-stage process and proposed the crossmodal 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.

Jaewoo 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.

Wenjun 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 of 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 imperious need for a novel method for addressing the complicated challenges associated with Audio-Enhanced Text-to-Video Retrieval.

3. 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 environment. 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

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

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 is 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.

4. 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.

Video Encoder and Text Encoder to Video-Text Cross Attention Block

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 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.

Audio Encoder and Text Encoder to Audio-Text Cross Attention Block

In addressing audio-related challenges, the Audio Spectrogram Short-Term Memory Transformer, the integration of Audio Spectrogram Transformer and Long Short-Term Memory (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 Temporal Graph Network (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, Dynamic Time Warping (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.

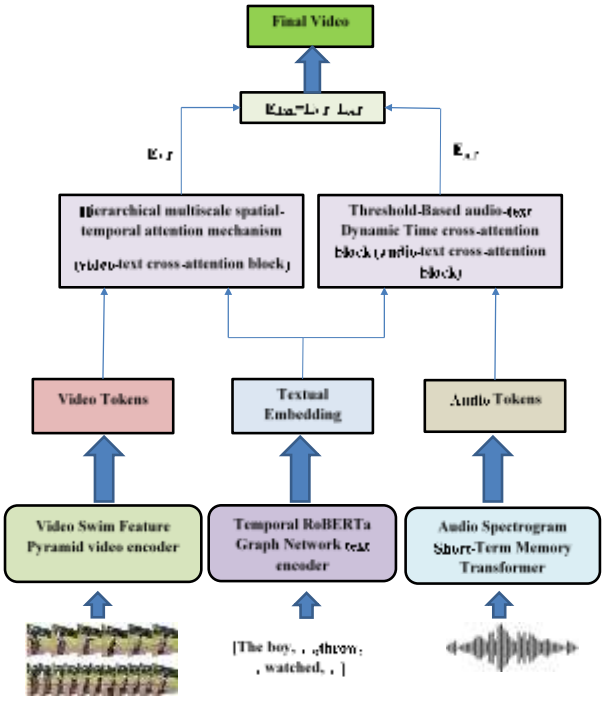


FIGURE 1. Diagram for the proposed method

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

The system integrates advanced video, audio, and text encoding mechanisms to achieve precise cross-modal alignment (figure 1). The Video Swim Feature Pyramid Network extracts multi-scale spatial features, enabling the Video-Text Cross Attention Block to align intricate video details with text embeddings derived from the Temporal RoBERTa Graph Network, which captures multi-theme relationships in the text. For audio, the Audio Spectrogram LSTM Transformer models temporal dynamics, isolating relevant sound events while the Threshold-Based Attention Mechanism filters out noise. Text embeddings, generated by the Temporal RoBERTa Graph Network, ensure robust contextual understanding. The Dynamic Time Warping (DTW) technique addresses temporal misalignments by stretching or compressing time, enhancing the Audio-Text Cross Attention Block for accurate alignment in complex acoustic environments. Together, these components resolve multi-scene transitions and ensure precise retrieval across video, audio, and text domains.

4.1. VIDEO SWIM FEATURE PYRAMID VIDEO ENCODER

A. Input to Video Encoder: Feature Extraction

The video input consists of a sequence of frames, $V = \{v_1, v_2, \dots, v_T\}$, 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 Feature Pyramid Network (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 $F = \{F_1, F_2, \dots, F_L\}$, where each F_l is a feature map at scale l, 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 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 $E_V = \{e_{v1}, e_{v2}, \dots, e_{vT}\}$, 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 equation (1).

$$E_V = FPN(v_1, v_2, \dots, v_T) \quad (1)$$

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.

B. Text Encoder: Temporal RoBERTa Graph Network

The The text input $T = \{t_1, t_2, \dots, t_N\}$, 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 $E_T = \{et_1, et_2, \dots, et_N\}$ that capture the semantic and syntactic relationships between the words.

However, the Temporal Graph Network (TGN) consider 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].

$$E_T = TGN\big(RoBERTa(t_1, t_2, \dots, t_N)\big) \quad (2)$$

Thus, (equation 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.

C. 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 equation (3):

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

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 $E_{VT} = \{e_{vt1}, e_{vt2}, \dots, e_{vtT}\}$, where each e_{vt_t} represents the aligned video-text token at time t , integrating both visual and textual information for each frame in equation (4).

$$E_{VT} = Cross - Attention(E_T, E_V) \quad (4)$$

D. 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 equation (5):

$$e_{CLS} = CLS(E_{VT}) \quad (5)$$

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.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

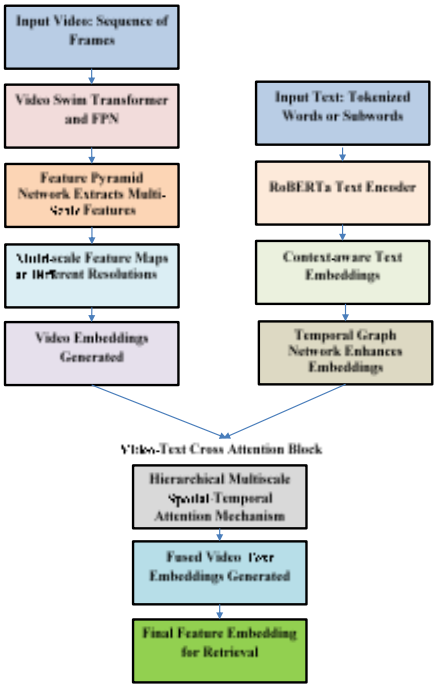


FIGURE 2. Flow diagram for Video Swim Feature Pyramid video encoder

The flow diagram (Figure 2) illustrates the process of aligning video and text features using a multi-stage approach. The video input, consisting of frames, is processed through a Video Swim Transformer and Feature Pyramid Network (FPN) to extract multi-scale features, generating embeddings for each frame. Simultaneously, the text input is tokenized and passed through a RoBERTa encoder, followed by a Temporal Graph Network (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 token (CLS) that condenses the aligned information for downstream tasks like retrieval.

4.2. AUDIO SPECTROGRAM SHORT-TERM MEMORY TRANSFORMER

A. 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 Audio Spectrogram Transformer (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 equation (6) [28]:

$$E_A^{spec} = \left\{e_{a_1^{spec}}, e_{a_2^{spec}}, \dots, e_{a_T^{spec}}\right\} \quad (6)$$

which encode the spectral features of the audio signal.

Next, these spectrogram embeddings are passed into a Long Short-Term Memory (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 equation (7):

$$E_A = \{e_{a1}, e_{a2}, \dots, e_{aT}\} \quad (7)$$

These final audio embeddings are enriched with both spectral and temporal information, making them suitable for cross-modal alignment with the text input in equation (8).

$$E_A = LSTM\big(AST(A)\big) \quad (8)$$

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. Text Encoder: Temporal RoBERTa Graph Network (Temporal and Contextual Embedding Generation)

The The text input $T = \{t_1, t_2, \dots, t_N\}$ 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 pretrained on large corpora using masked language modeling. RoBERTa applies self-attention over the text tokens to generate contextually-rich embeddings in equation (9) [29]:

$$E_T^{ctx} = \{e_{t1}^{ctx}, e_{t2}^{ctx}, \dots, e_{tN}^{ctx}\} \quad (9)$$

where each embedding captures the semantic relationships between the words in the sentence.

To further enhance the temporal understanding of the text, the Temporal Graph Network (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: $E_T = \{e_{t1}, e_{t2}, \dots, e_{tN}\}$, where each embedding is enriched with both semantic context and

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

temporal structure. These text embeddings are prepared for cross-attention with the audio embeddings in the subsequent block in equation (11).

$$E_T = TGN\big(RoBERTa(T)\big) \tag{10}$$

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.

C. 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} and each audio token e_{at}. 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 equation (11):

$$\alpha_{t,a} = \frac{\exp(\langle e_{tn}, e_{a_t} \rangle)}{_k{}^{\exp(\langle e_{tn}, e_{a_k} \rangle)}} \tag{11}$$

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, Dynamic Time Warping (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 equation (12):

$$E_{AT} = \{e_{at1}, e_{at2}, \dots, e_{atT}\} \tag{12}$$

where each token contains information from both the audio and text modalities, effectively aligned for further processing in equation (13).

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

The output of the Audio-Text Cross-Attention Block, denoted as $E_{AT} = \{e_{at1}, e_{at2}, \dots, e_{atT}\}$ 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 $A = \{a_1, a_2, \dots, a_T\}$
- Text sequence $T = \{t_1, t_2, \dots, t_N\}$

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 $E_A^{spec} = \left\{e_{a_1^{spec}}, e_{a_2^{spec}}, \dots, e_{a_T^{spec}}\right\}$
3. Pass to Long Short-Term Memory (LSTM):
   - Input: $E_A^{spec} = \left\{e_{a_1^{spec}}, e_{a_2^{spec}}, \dots, e_{a_T^{spec}}\right\}$
   - Process: LSTM learns temporal dependencies.
   - Output: Contextually-aware audio embeddings $E_A = \{e_{a1}, e_{a2}, \dots, e_{aT}\}$
   - Formula: $E_A = LSTM\big(AST(A)\big)$

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 $E_T^{ctx} = \{e_{t1}^{ctx}, e_{t2}^{ctx}, \dots, e_{tN}^{ctx}\}$
3. Pass to Temporal Graph Network (TGN):
   - Input: E_T^{ctx}
   - Process: Construct a graph-based representation for temporal relationships.
   - Output: Text embeddings $E_T = \{e_{t1}, e_{t2}, \dots, e_{tN}\}$
   - Formula: $E_T = TGN\big(RoBERTa(T)\big)$

Step 3: Audio-Text Cross-Attention
1. Calculate Attention Weights:
   - Input: e_{at}, e_{tn}
   - Process: Compute attention scores $\alpha_{t,a} = \frac{\exp(\langle e_{t_n}, e_{a_t} \rangle)}{k^{\exp(\langle e_{t_n}, e_{a_k} \rangle)}}$
2. Apply Threshold-Based Mechanism:

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

- 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: $E_{AT} = \{e_{at1}, e_{at2}, \dots, e_{atT}\}$
   - Formula: $E_{AT} = DTW\big(Threshold - Based\ Cross - Attention(E_T, E_A)\big)$

Final Outputs:

- Audio Embeddings: E_A – Contextually enriched audio tokens.
- 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 Short-Time Fourier Transform, followed by the application of the Audio Spectrogram Transformer (AST) to extract spectral features. These features are then passed through a Long Short-Term Memory (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 Temporal Graph Network (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 Dynamic Time Warping 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.

4.1. 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 $A = \{a_1, a_2, \dots, a_T\}$ 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 equation (14):

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

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 equation (15).

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

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 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.

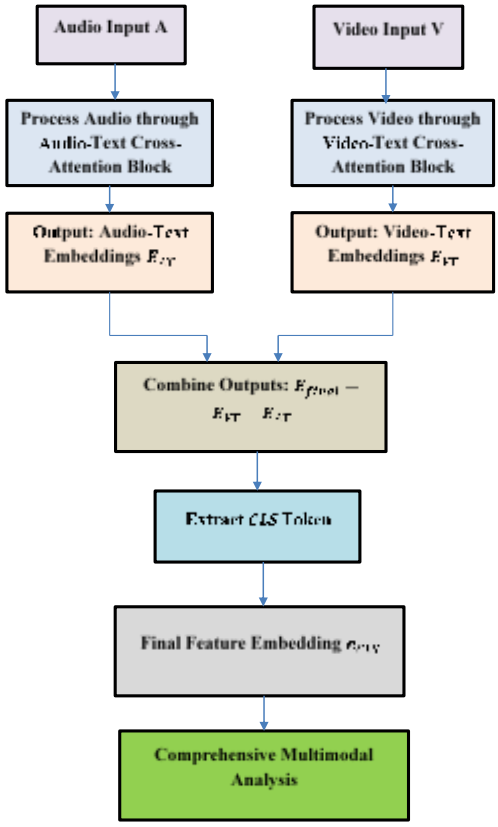


FIGURE 3. Final Summation Output in Multimodal Systems

The flowchart begins with the input data, where audio and video inputs are represented as "Audio Input A" and "Video Input V," respectively (Figure 3). 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-

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

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.

5. RESULTS AND DISCUSSION

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

5.1. TOOLS AND SPECIFICATIONS

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

5.2. PERFORMANCE OF THE PROPOSED METHODS

FIGURE 4. Heat Map

This heatmap represents attention weights between text tokens (Token 1 to Token 5) and video segments (Segment 1 to Segment 10), where darker shades indicate higher attention weights (Figure 4). The model dynamically assigns different levels of importance to various text-video pairs. For instance, Token 4 shows strong attention alignment with Segment 5 (0.97) and Segment 4 (0.95), indicating that these video segments closely relate to the context of this specific token. 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. By adjusting the attention dynamically based on contextual relevance, it improves the alignment of fine-grained video features with

textual semantics, offering better retrieval precision, especially for videos with complex scene transitions and varying themes. This novel mechanism enhances performance by capturing the nuanced relationships between segments and tokens, leading to more accurate cross-modal alignments.

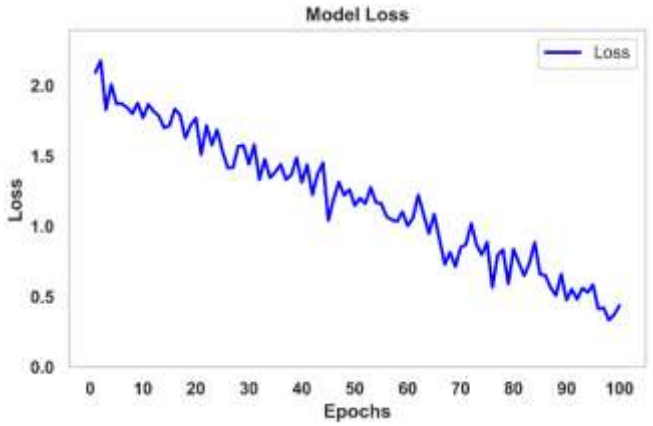


FIGURE 5. Model loss

This graph represents the model loss over 100 epochs, showing a clear downward trend, indicating that the model's performance improves as training progresses (Figure 5). The initial loss is around 2.0, and it steadily decreases to below 0.5, suggesting effective learning. Fluctuations early on are common during the initial phases of optimization, but the consistent decline suggests that the model is converging effectively. The Hierarchical Multiscale Spatial-Temporal Attention Mechanism (from your prior descriptions) improves the alignment of complex, cross-modal data (e.g., video-text or audio-text), which accelerates convergence by focusing on the most relevant features and filtering noise, leading to faster and more stable reduction in loss over time. This results in better model generalization and accuracy.

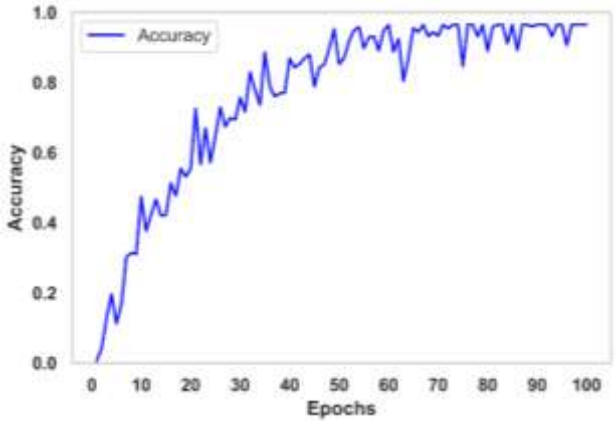


FIGURE 6. Accuracy

This graph shows the accuracy of the model improving over 100 epochs (Figure 6). Initially, the accuracy is very low, close to 0, but it increases rapidly, reaching near 0.9 by epoch 50, and stabilizes at around 1.0 as training progresses. There are minor fluctuations, but the overall trend shows consistent improvement, indicating strong model convergence. The Threshold-Based Dynamic Attention Mechanism enhance accuracy by dynamically focusing on the most relevant parts of the input data, be it video, audio, or text, and reducing noise or irrelevant segments. By fine-tuning the

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

cross-attention across multiple scales and adjusting to temporal variations, the model ensures better feature alignment, which leads to more precise predictions, reflected in the high accuracy achieved.

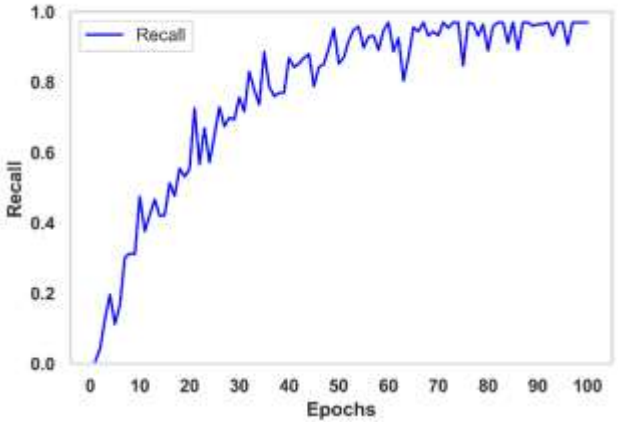


FIGURE 7. Recall

This graph shows the recall of the model improving over 100 epochs (Figure 7). Initially, the recall is very low, close to 0, but it increases rapidly, reaching near 0.8 by epoch 50, and stabilizes at around 1.0 as training progresses. There are minor fluctuations, but the overall trend shows consistent improvement, indicating strong model convergence. The novel Threshold-Based Dynamic Attention Mechanism enhance recall by dynamically focusing on the most relevant parts of the input data, whether video, audio, or text, and filtering out noise or irrelevant components. By fine-tuning the cross-attention across multiple scales and adapting to temporal variations, the model ensures better feature alignment, leading to more accurate and comprehensive detection of relevant instances, reflected in the high recall achieved.

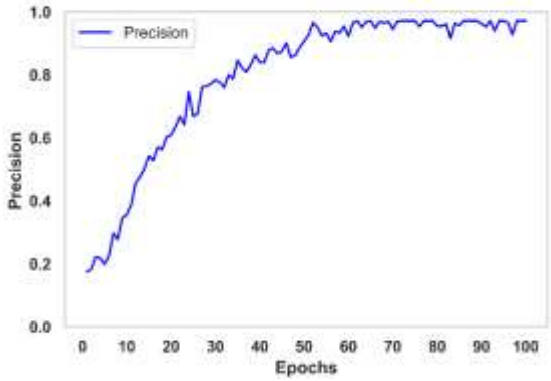


FIGURE 8. Precision

This graph (Figure 8) shows the precision of the model improving over 100 epochs. Initially, the precision is low, close to 0.2, but it gradually increases, reaching approximately 0.9 by epoch 50 and stabilizing near 1.0 as the training progresses. There are some minor fluctuations, but the general trend shows a consistent improvement, suggesting strong model performance and convergence. The novel Hierarchical Multiscale Spatial-Temporal Attention Mechanism enhance precision by dynamically filtering out irrelevant or noisy input data, whether video, audio, or text,

and focusing only on the most relevant elements. By leveraging multiscale attention and adjusting focus based on temporal structures, the model aligns critical features effectively, minimizing false positives and thereby improving precision as seen in the graph.

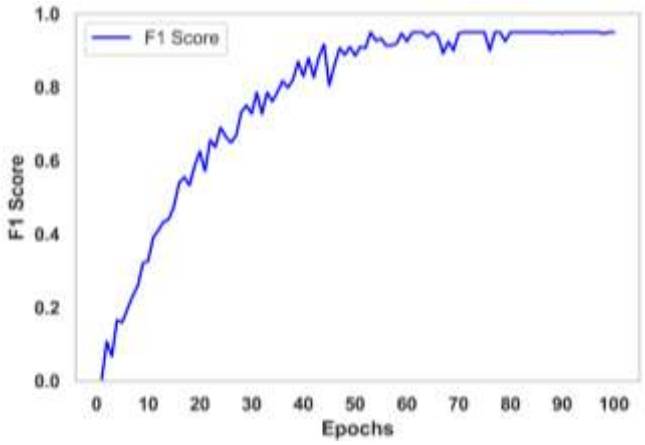


FIGURE 9. F1Score

This graph (Figure 9) illustrates the F1 score of the model improving over 100 epochs. Initially, the F1 score is low, close to 0, but it rapidly increases, reaching around 0.9 by epoch 50 and stabilizing near 1.0 as training progresses. Minor fluctuations are present, but the overall trend shows consistent improvement, indicating a balanced trade-off between precision and recall. The novel the Adaptive Cross-Layer Attention technique enhance the F1 score by improving the model's ability to balance precision and recall through dynamic attention shifts across different temporal and spatial features. By effectively managing the relationship between relevant data and minimizing false positives and false negatives, these attention mechanisms help achieve the high F1 score reflected in the graph.

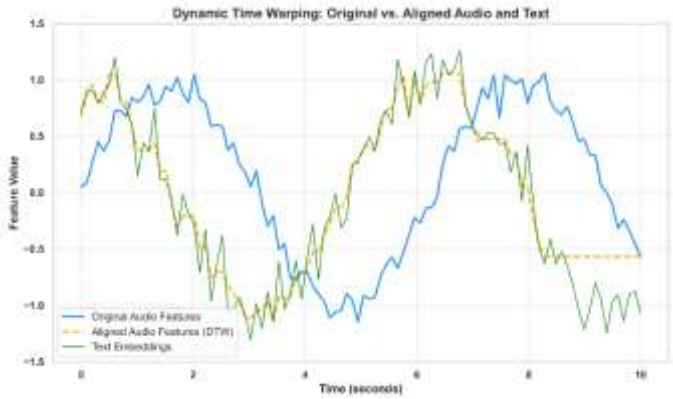


FIGURE 10. Feature Values of DTW

This graph (Figure 10) compares original audio features (blue line) with aligned audio features using Dynamic Time Warping (DTW) (orange dashed line) and text embeddings (green line). The purpose of the graph is to showcase how DTW helps align audio features with text embeddings over time, addressing misalignments due to

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

temporal variations. The novel method here is the Threshold-Based Audio-Text Dynamic Time Cross-Attention, which incorporates DTW to handle time stretching and compression in audio signals. By aligning temporally dispersed audio features more closely with text embeddings, this approach improves retrieval accuracy. The method ensures precise alignment even in noisy or asynchronous audio environments, which leads to better cross-modal attention and retrieval.

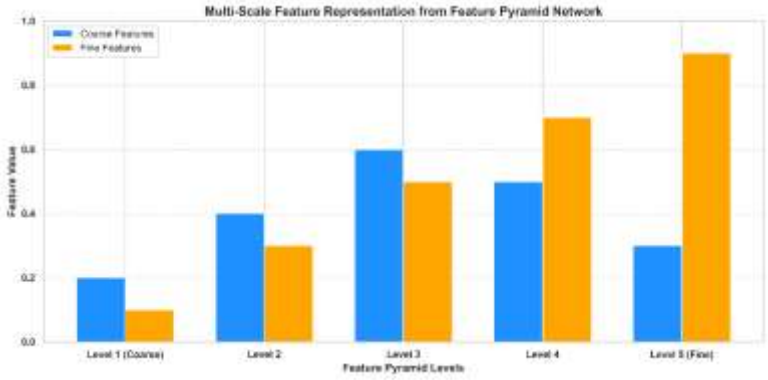


FIGURE 11. Feature Values of FPN

This graph (Figure 11) illustrates the Multi-Scale Feature Representation from a Feature Pyramid Network (FPN) across five levels, showing coarse features (blue) and fine features (orange). As the levels progress from coarse (Level 1) to fine (Level 5), fine feature representation increases, while coarse feature values remain steady or decrease. The novel method related to this is the Video Swim Transformer integrated with FPN, which extracts multi-scale features. By capturing both coarse and fine-grained details, the FPN enables more accurate alignment between video segments and text queries. The hierarchical multiscale attention mechanism takes advantage of these diverse feature levels, improving the model's ability to focus on both high-level context and detailed scene transitions, leading to more precise retrieval and alignment in video-text tasks.

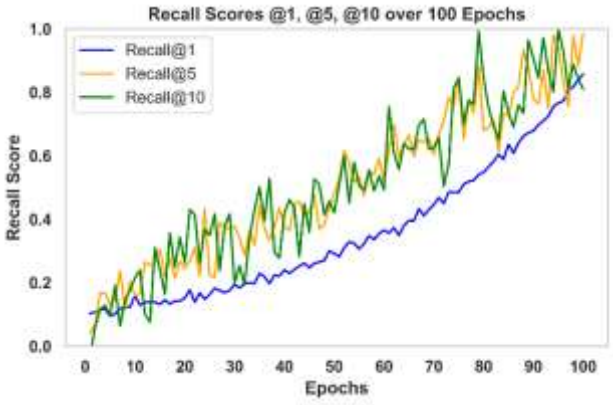
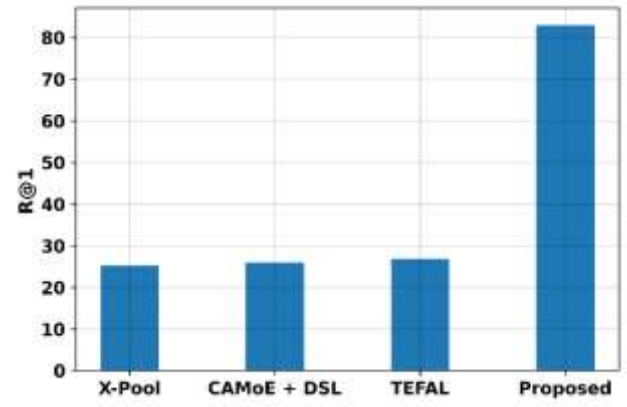


FIGURE 12. Recall score (@1, @5, @10)

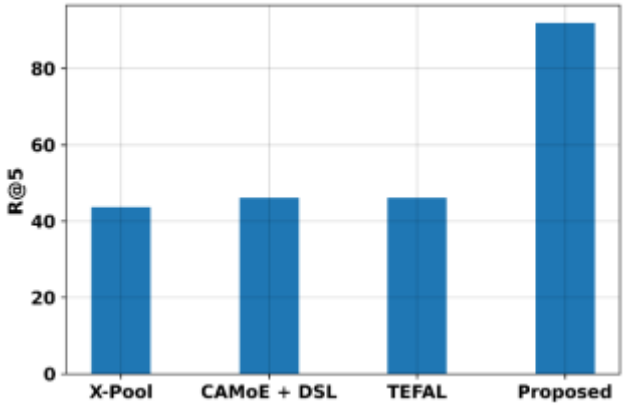
This graph (Figure 12) shows Recall scores at different thresholds (@1, @5, @10) over 100 training epochs. As the epochs progress, all recall scores improve, with Recall@10 (green) achieving the highest performance, followed by Recall@5 (orange), and Recall@1 (blue). The novel method here is likely the hierarchical

multiscale spatial-temporal attention combined with Dynamic Time Warping (DTW) for video and audio-text alignment. This method dynamically refines attention across multiple temporal and spatial scales, improving the model's ability to retrieve the most relevant video or audio segments. The increasing recall scores across epochs demonstrate the effectiveness of this approach, particularly in handling complex, multi-scene data and misaligned audio-text signals, leading to more accurate retrieval results as training continues.

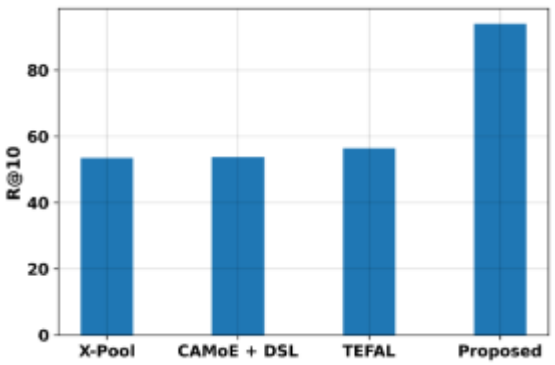
5.3. PERFORMANCE OF THE PROPOSED METHODS



(a)



(b)



(c)

FIGURE 13. Recall (Suggested versus Existing Method)

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

a) R@1: The first graph illustrates that the proposed model significantly outperforms the other models, 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.
b) R@5: In the second graph, 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.
c) R@10: The final graph 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.

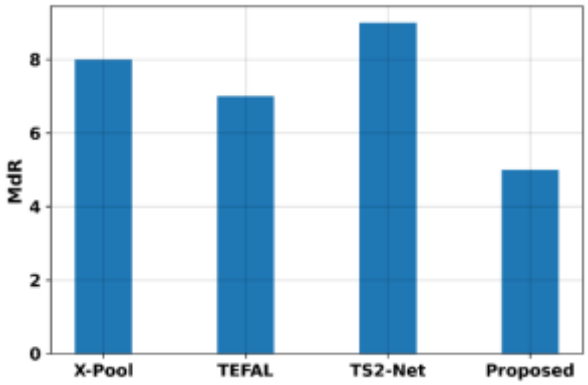


FIGURE 14. MdR (Suggested versus Existing Method)

The graph (figure 14) compares the performance of four models—X-Pool, TEFAL, TS2-Net [31], and the Proposed model—using the MdR (Median Rank) metric. 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.

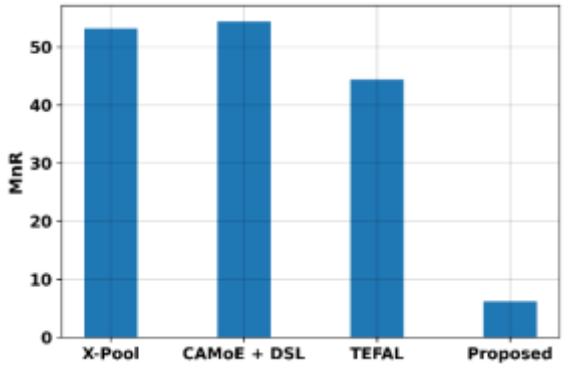


FIGURE 15. MnR (Suggested versus Existing Method)

The graph (figure 15) compares the performance of four models—X-Pool, CAMoE + DSL, TEFAL [31], and the Proposed model—based on the MnR (Mean Rank)

metric. X-Pool and CAMoE + DSL both have the highest MnR values around 50, indicating worse performance in ranking tasks. TEFAL performs slightly better with an 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 mean rank by a substantial margin.

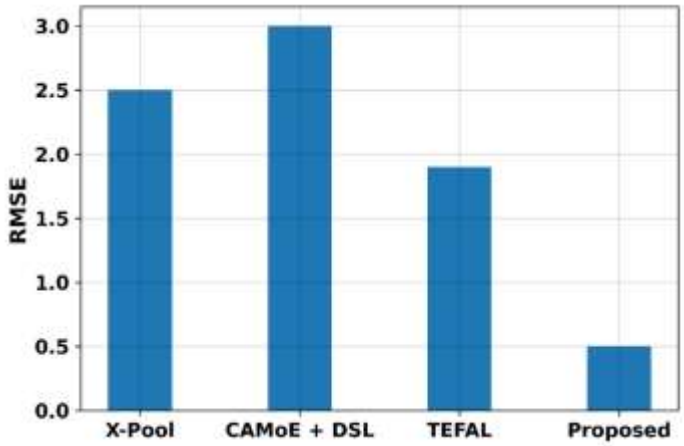


FIGURE 16. RMSE (Suggested versus Existing Method)

The graph (figure 16) presents a comparison of the Root Mean Square Error (RMSE) values across four different models: X-Pool, CAMoE + DSL, TEFAL, and 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**.**

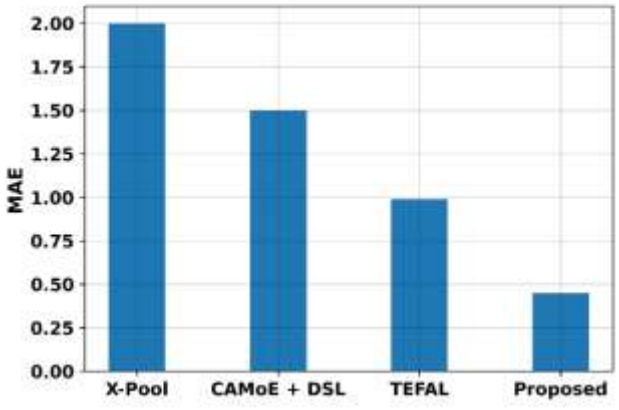


FIGURE 17. MAE (Suggested versus Existing Method)

The graph (Figure 17) compares the Mean Absolute Error (MAE) values of four different models: X-Pool, CAMoE + DSL, TEFAL, and the Proposed model. The bars represent the MAE, indicating that the Proposed model achieves the lowest MAE at 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.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Overall, the graph emphasizes the effectiveness of the Proposed model in minimizing error compared to existing methodologies.

Overall, the suggested 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 like X-Pool, CAMoE + DSL, and TEFAL show significantly lower recall rates, highlighting their limitations. The Median Rank (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 Mean Rank (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 Root Mean Square Error (RMSE) of 0.5, outperforming CAMoE + DSL, X-Pool, and TEFAL. The Mean Absolute Error (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.

tables should be consecutively numbered, and should be horizontally centered across the page, with their headings (in Times New Roman font, point size 10) placed centered above the table. Each table should be placed as close as possible to where the table is first mentioned in the text. All text in tables should be in Times New Roman font, in either point sizes 8 or 9. Refer to Table 1 for guidance.

6. 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 Feature Pyramid Network (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 Audio Spectrogram Transformer combined with Long Short-Term Memory (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 Median Rank (MdR) and Mean Rank (MnR), as well as superior performance in reducing Root Mean Square Error (RMSE) to 0.5 and Mean Absolute Error (MAE) to 0.25, outperforming existing methods such as X-Pool, TEFAL, and CAMoE + DSL. This comprehensive approach improves retrieval accuracy, precision, and ranking effectiveness across diverse video and audio content.

REFERENCES

[1] 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*.

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

[3] 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*.

[4] 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).

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

[6] 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.

[7] 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*.

[8] 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*.

[9] 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.

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

[11] 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.

[12] 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*.

[13] 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*.

[14] 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*.

[15] 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*.

[16] 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

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 7, pp. 6639-6647).

[17] 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.

[18] 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).

[19] 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).

[20] 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.

[21] 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).

[22] 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.

[23] 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).

[24] 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*.

[25] 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*.

[26] 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).

[27] 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).

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

[29] 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.

[30] 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.

[31] 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).

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)