Cross-Modal Video Retrieval Model Based on Video-Text Dual Alignment

Zhanbin Che, Huaili Guo*
College of Computer, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China

***Abstract*—Cross-modal video retrieval remains a major challenge in natural language processing due to the natural semantic divide between video and text. Most approaches use a single encoder to extract video and text features separately, and train video-text pairs by means of contrastive learning, but this global alignment of video and text is prone to neglecting more fine-grained features of both. In addition, some studies focus only on profiling the video description text, ignoring the correlation relationship with the video. Therefore, this paper proposes a video retrieval method based on video-text alignment, which realizes both global and fine-grained alignment between video and text. For global alignment, the video and text are aligned by a single encoder and after linear projection; for fine-grained alignment, the video encoder is trained to align the video and text by masking some semantic information in the text. By experimentally comparing with multiple existing methods on MSR-VTT and MSVD datasets, the model achieves R@1 (recall at 1) metrics of 51.5% and 52.4% on MSR-VTT and MSVD datasets, respectively, which indicates that the proposed model can improve the efficiency of cross-modal video retrieval.**

Keywords*—*Video-text alignment; cross-modal; contrastive learning; similarity measure; feature fusion

I. INTRODUCTION

With the proliferation of mobile devices and high-speed networks, network resources predominantly manifest in textual and video formats. Video's formidable capacity for conveying information confers upon it a distinct advantage, rendering it more popular among users. Video retrieval not only reduces costs but also fosters innovation, enhances the quality of life, and generates economic value in diverse fields such as education, military, and healthcare. Consequently, the demand for precision in video content retrieval is escalating, making the enhancement of video retrieval efficiency a formidable research pursuit. Within the realm of video comprehension, a natural semantic gap exists between the various modalities of video. Solely extracting semantic features from videos is susceptible to yielding sparse feature representations, consequently diminishing the accuracy of video retrieval. Consequently, numerous scholars have endeavored to represent video features through multiple modalities to augment the precision of video retrieval, yielding noteworthy results. Current models such as Frozen [1], CLIP4Clip [2], and Clipbert [3] use contrast learning to achieve semantic alignment and interaction of cross-modal features, where features from different modalities are extracted and then mapped into the same space, enabling global alignment of video with video description text.

The semantic alignment strategy for unimodal encoders in comparative learning typically involves the integration of features from video description text and video features to calculate their similarity. However, this approach often overlooks the association between the local features of the two modalities, resulting in asymmetry in their representation and impacting the efficiency of cross-modal retrieval. In addressing these issues, some researchers employ lexical embedding [4] to achieve fine-grained retrieval by leveraging the relationship between different lexemes. Chen et al. [5] introduced a hierarchical graph inference model to generate text embeddings using an attention-based graph inference mechanism, capturing global-to-local feature associations. Notably, this model primarily focuses on text comprehension and neglects alignment with video content. To address this limitation, HANet [6] enhances the alignment between video and text by introducing a word-level attention mechanism. This mechanism calculates the importance of each word in the video representation and weights the text representation accordingly. However, the computational complexity of HANet is high due to the incorporation of a multi-level attention mechanism.

Upon a thorough examination of relevant research, it becomes evident that video retrieval models should prioritize video sub-regions closely associated with a given video summary. This entails employing cross-modal reasoning between video summaries and video frames to identify the most semantically relevant segments in both, thereby achieving alignment between the video and the summary text. However, prevailing video retrieval models frequently rely on global features of videos, utilizing mean pooling or self-attention methods. Unfortunately, these approaches fall short in effectively integrating the concept of cross-modal reasoning in practical applications. Consequently, the lack of fine-grained semantic attention to both video and summary text within the global alignment model hinders the encoding of localized visual information in the video. This deficiency subsequently leads to a degradation in retrieval performance.

In this paper, we introduce a dual video-text alignment model that aims to narrow the semantic gap between video and text at a finer granularity, thus improving the efficiency of video retrieval. Initially, a conventional methodology is employed to map features from both the video and the summary text into a shared space. This facilitates the computation of contrast loss, thereby achieving global alignment between the video and the text. Subsequently, we concentrate on the actions or scenes involving entities in the video, aligning them with the nouns and verbs present in the

*Corresponding Author.

textual description. This dual-pronged approach not only establishes global alignment but also enables more refined local alignment. The result is a comprehensive interaction between video and text, enhancing retrieval performance.

The rest of this paper is as follows, Section II review previous studies. Section III discusses the methodology. Section IV presents experimental setup. Section V describes the results of the experiment and discusses. Finally, conclusion presents in Section VI.

II. Related Works

This section provides an overview of related work on video retrieval methods and the video-text alignment method used in this paper, where exploring a new video retrieval method is the target task of this paper, and the study of the video-text alignment method is the focus of this paper.

A. Video Search Methods

Video-text alignment methods are more commonly used in video retrieval tasks. In their earlier work, Kaufman et al. [7] focused on pre-training by designing cross-modal fusion mechanisms, utilizing large-scale multimodal data for pre-training, and fine-tuning in downstream tasks. However, these approaches usually focus only on the global alignment of video and text, ignoring the interaction of local representations and affecting the retrieval efficiency.

Currently, popular methods encode video and text into feature vectors that are projected into a common space for matching by means of a dual-encoder structure of a text encoder and a video encoder. These methods utilize dot product operations to compute the global similarity and thus achieve alignment between video and text. For example, Bain et al. [1] proposed an end-to-end model that utilizes the ideas of ViT [8] and Transformer [9] to achieve a common representation of video and text. Luo et al. [2] utilized the knowledge migration of CLIP [10] to match the feature vectors of the video and the text in the common space and retrieve them by using the similarity between the vectors. Li et al. [11] matched multiple encoders in a specific common space, avoiding the dominant role of a single encoder and failing to fully utilize the visual information within the video, relying too much on textual information. In addition, methods based on graph neural networks [12], which represent video and text as graph structures; utilize graph neural networks for information dissemination and fusion. However, these methods still have some problems:

1) Global similarity may not adequately capture the complex relationships between video and text. Since video and text have different structures and semantics, relying only on global similarity may ignore the interaction of local representations.

2) Inconsistency in the length of video and text may lead to information loss. In a dual-encoder architecture, the output of the text encoder is usually truncated to fit the input of the video encoder. This truncation may lead to loss of textual information during the encoding process, thus affecting retrieval.

3) Imbalance of training data may lead to model overfitting. In video-text retrieval tasks, the training data is usually unbalanced, which may lead to overfitting of the model to local similarities during the training process, while ignoring the importance of global similarities.

To address these problems, researchers have proposed some improvement strategies, such as introducing an attention mechanism and utilizing methods such as contrast learning to capture local and global representations between video and text, which can effectively improve the performance of video-text retrieval tasks.

B. Video-Text Alignment Methods

Video-text alignment is commonly used in application domains such as video retrieval, video annotation, and video quizzing, and using features and objects in the video to match with the text is a common alignment method. Some work relies on the attention mechanism to extract information from videos [13], which is then used in downstream tasks such as video quizzing. Wang et al. [14] utilized multiple pre-trained experts to extract multimodal information and use it as an anchor point for alignment with text. Dong et al. [15] designed dual coding networks to perform multilevel coding of video spatial and temporal information with text. Luo et al. [2] based on the inspiration of a large-scale pre-trained graphic-text matching model CLIP [10], migrated the image-text alignment method to video-text alignment to realize video text retrieval. However, all these alignment methods are global alignment on the whole of video and video description text, ignoring the finer-grained semantic information alignment between the two.

To solve the above problems, some works split text descriptions into semantic phrases, e.g., Yang et al. [16] construct a semantic tree representation of the text and use a temporal attention encoder to obtain a video representation. Wang et al. [17] manipulate a fine-grained comparison target by selecting video frames that are semantically equivalent to the text to better learn the representation of the video and the text. Chen et al. [5] extract the text from the sentence to extract verbs and nouns and project them into a shared space for fine-grained alignment. In addition, Li et al. [18] performed large-scale image-text comparison learning through the Twin Towers model, which aligns the visual information in images with the meaning of text masked words through Masked Language Model (MLM) and Image-Text Matching (ITM) to achieve image-text retrieval.

Synthesizing the research on video retrieval algorithms, this paper proposes effective solutions to the problems of the current methods in Section *A*, and the main contributions include:

1) A video-text dual alignment model is proposed to enhance the interaction of local representations by aligning the global and fine-grained features of video with those of text to capture the more complex semantic relationships between the two.

2) Using the Transformer-based dual encoder structure, text information can be encoded more comprehensively, reducing the missing information caused by truncated text features.

3) Using the contrast learning method, video features are fitted to sentence features with global and fine-grained similarity to increase the balance of the training data and improve the video retrieval accuracy.

III. METHODOLOGY

The video-text dual alignment method employs dual encoders to train video and text features, proposing a dual alignment model for both global and fine-grained alignment to enhance cross-modal alignment between videos and summary text for improved video retrieval accuracy.

In this section, Section *A* provides an overview of the model to illustrate its working principles and processes, Section *B* details how video and text features are extracted, Section *C* introduces the framework and working principles of the dual alignment network model, Section *D* outlines the model's training strategy. Finally, Sections *E* and *F* elaborate on the objective functions and pretraining datasets used in this approach.

A. Overview of the Model

As illustrated in Fig. 1, our model adopts a two-tower structure to capture semantic information from both video and text during the feature extraction phase, employing dedicated encoders for each modality. For a given set of videos, the TimeSformer [19] serves as the video feature encoder. The output features, denoted as x_i and v_i, plays crucial roles in the global and fine-grained alignment of video and text, respectively. In the case of video description text, two inputs are provided to the text encoder DistilBERT [20]: the sentence with deleted verbs and nouns, and the complete video text description. The extracted text features y_i and t_j are used for global alignment and fine-grained alignment, respectively.

During the video-text alignment phase, one branch focuses on global alignment, projecting global video and text features into a common space. This branch is trained using a contrastive learning method to attain comprehensive semantic alignment of features on a global scale. The other branch is dedicated to fine-grained alignment, achieving detailed alignment between video and text by training a multimodal encoder to acquire vector representations of deleted nouns and verbs in the video modality.

The model enhances its cross-modal alignment capability, specifically in video-text alignment, through the incorporation of a cross-modal attention mechanism. This mechanism projects both the video and text into an embedding space, where semantic similarity is maximized. Consequently, for a given text query, video retrieval is formulated as a cross-modal similarity metric, aiming to identify videos that exhibit semantic alignment with the query.

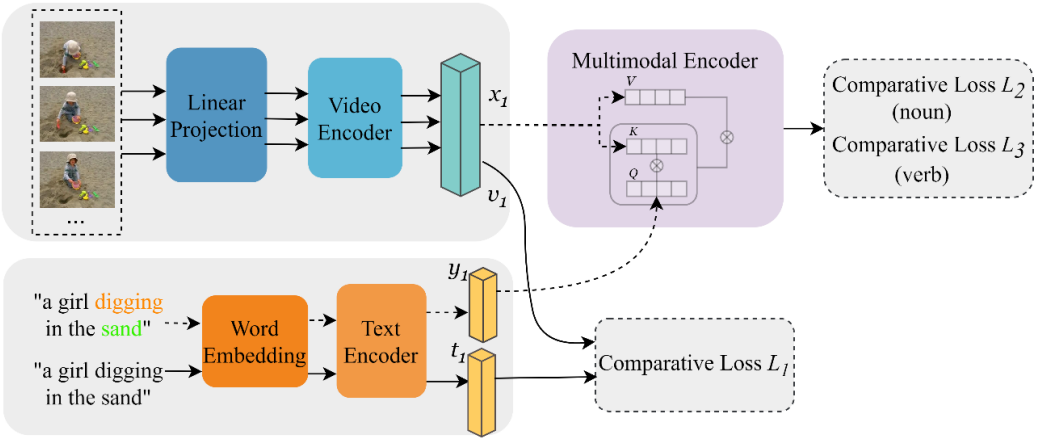


Fig. 1. Video-text dual alignment framework.

B. Feature Extraction

In this approach, both video and summary text undergo parsing using a video encoder and a text encoder. The distinct data modalities are then transformed into a unified numerical representation, yielding a sequence of feature representations for both. This facilitates alignment between the video and summary text on both a global and fine-grained level.

1) Video representation: This paper employs a dual encoder architecture for the extraction of video and text features. Specifically, TimeSformer is utilized to extract video features for each video. During the extraction of video features, as illustrated in Fig. 1, the M video frames of the clip are initially input into TimeSformer. Each video frame is then partitioned into P patches, which are subsequently fed into a linear projection header, spreading them into a series of tokens for the video. Following this, learnable $[CLS]$ tokens are affixed to the header of the sequence, enhancing our ability to learn sentence-level features for downstream tasks. Learnable positional embeddings are also introduced to the tokens. For each video frame feature $v_i \in R^{M\times P\times D}$, with D representing the feature dimension, TimeSformer applies the self-attention mechanism in both temporal and spatial dimensions, generating the final sequence of video frame embeddings $v_i = \{v_{cls}, v_1, v_2, ..., v_p\}$.

2) Text representation: For each of the N text descriptions associated with each video, this approach employs the DistilBERT model for feature extraction.

DistilBERT, being a lightweight BERT [21] model, is more suitable for deployment and operation under resource constraints due to its smaller size compared to the BERT model. DistilBERT produces a text embedding sequence, denoted as $t_j \in R^{N\times D}$, by tagging text description embeddings $[CLS]$ and positional tags, resulting in text features represented by $t_j = \{t_{cls}, t_1, t_2, ..., t_N\}$.

C. Model Framework

This paper centers on the examination of video-text alignment methods, emphasizing the double alignment of feature representations for both video and descriptive text at both global and fine-grained levels. This approach aims to enhance the overall understanding of the video and descriptive text, thereby improving the retrieval accuracy of the model.

1) Global alignment: In the context of global comparative learning, for a given video-text pair, subsequent to extracting features using two distinct encoders, the video embedding sequences and text embedding sequences are initially projected into a shared space through linear projection. Subsequently, all frames of each video undergo aggregation using mean pooling to obtain the average frame $\overline{v_i}$, $\overline{v_i} \in R^{1\times D}$. For each text, this paper extracts the representation by taking the first $[CLS]$ token, denoted as $\overline{t_j}$, $\overline{t_j} \in R^{1\times D}$. Finally, the method computes the similarity between them using the cosine similarity function denoted by $s(v_i, t_j)$. During training, the objective is to maximize the correct pairing of video-text pair comparisons while minimizing the remaining comparison targets that cannot be paired. The cosine similarity function is defined as shown in Eq. (1).

$$s(v_i, t_j) = \frac{\overline{v_i} \bullet \overline{t_j}^{T}}{\left\|\overline{v_i}\right\| \left\|\overline{t_j}\right\|} \tag{1}$$

2) Fine-grained alignment: In the domain of video-text alignment, when given a video and its corresponding text description, this approach involves the removal of nouns from the text, utilizing the incomplete sentence with omitted nouns as the text to be aligned. The sentence then undergoes processing through a text encoder to obtain an intermediate text sequence representation $\{n\}_{n\_t}$. Simultaneously, the video is processed through a video encoder to acquire the intermediate video sequence representation $\{c\}_v$.

Subsequently, the linear transformation of the noun text sequence $\{n\}_{n\_t}$ is treated as the query (Q), and the linear transformation of the video sequence $\{c\}_v$ serves as keys and values (K,V). Through cross-modal attention using the Transformer, these are interacted to obtain vector representations of nouns that can be aligned in the video modal space. The nouns, removed from the text, are also processed through the text encoder to obtain a text space vector representation of the nouns, which are used to form positive and negative samples in the comparison target.

The video sequence representation and the noun sequence representation are projected through two separate linear layers into a common embedding space, and their similarity is computed using a cosine similarity function. The formulation of the cross-modal attention mechanism is depicted in Eq. (2):

$$Attention(Q,K,V) = Soft\max(\frac{QK^{T}}{\sqrt{d_k}})V \tag{2}$$

where, Q denotes the text sequence originating from the deletion of the noun in the sentence, K and V denote the sequences originating from the video representation, and d_k denotes the dimension of K.

Similarly, the same operation is executed when removing a verb from a text as when omitting a noun. The sequence representation $\{v\}_{v\_t}$ of the sentence lacking the verb is derived, linearly transformed as a query (Q) in cross-modal attention, and interacted with the linear transformation of the video sequence representation $\{c\}_v$ to obtain the vector representation of the verb in the video modal space that can be aligned.

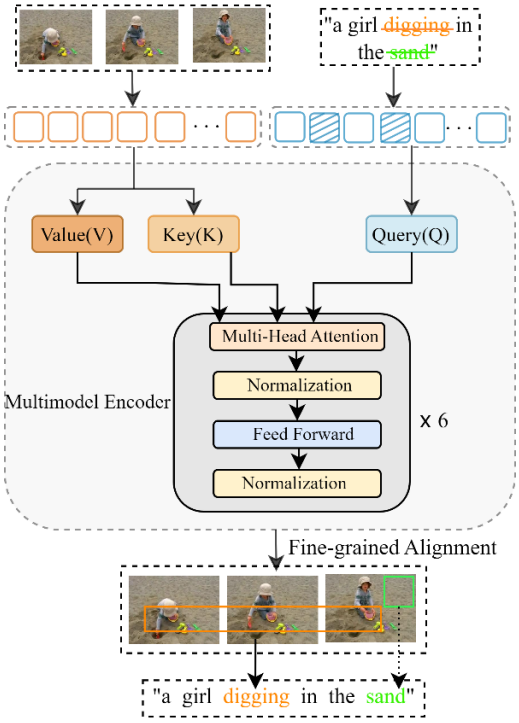


Fig. 2. Fine-grained alignment process.

The removed verbs also undergo processing through a text encoder to obtain a text space vector representation of the verb. The similarity is computed after passing these two representations through separate linear layers.

The fine-grained alignment process is depicted in Fig. 2, where this model linearly transforms the two modal features extracted from the video and the text, each with the noun or verb removed. The linearly transformed Q, K, and V are

employed as inputs to the multimodal encoder, which captures cross-modal attentions between the video and the text through the multi-head attention module.

Following this, a contrast learning approach is utilized to maximize the similarity of correctly paired nouns or verbs and minimize the vice versa scenario. The training model is adept at extracting semantic information, such as scenes or actions, from the video content that aligns with the nouns or verbs in the text. This enables a higher degree of fine-grained alignment between the video and the text, consequently enhancing the efficiency of video retrieval.

D. Training Strategy

In the contrast learning model presented in this paper, it projects input samples into a low-dimensional vector space. The model undergoes initial training, ensuring that similar samples in the vector space are mapped to proximate locations, while dissimilar samples are mapped to distant locations. Specifically, for an input sample x_i and a positive sample y_i, the model is trained by maximizing their similarity. Simultaneously, within the same batch, unpaired samples are considered as negative samples. The objective is for the model not only to match positive samples but also to distinguish them from negative samples.

Throughout the training process, for the given video-text pairs, this paper employs comparison learning utilizing the loss function based on cosine similarity scores, as demonstrated in Eq. (3):

$$L_{nce} = -\sum_{i=1}^{N} log \frac{exp(\, s(x_i, y_i) / \tau \,)}{\sum_{i=1}^{N} exp((x_i, y_j) / \tau \,)} \tag{3}$$

where, $s(\, x_i, y_i \,)$ denotes the similarity score between input samples x_i and samples y_i, and N represents the batch size. The temperature coefficient τ is a hyperparameter that must be set to control how effectively the model discriminates between negative samples. The essence of this loss function lies in the fact that for each sample x_i, this paper normalizes its similarity score with the positive sample y_i by dividing it by the sum of the similarity scores between x_i and all the samples. This process yields a probability distribution, and the logarithm of this distribution is incorporated into the loss function, which is then averaged across all sample results. The objective of this loss function is to maximize the similarity of positive samples while minimizing the similarity with negative samples, aiming to learn a comprehensive feature representation.

In the global alignment comparison learning of video and text, the objective is to maximize the similarity of correctly paired video-text pairs and minimize the similarity of those that cannot be paired. Subsequently, the video and text representations $\overline{v}_i$ and $\overline{t}_j$ outlined, the contrast loss is computed as expressed in Eq. (4):

$$L_1 = -\sum_{i=1}^{N} log \frac{exp(\, s(\overline{\boldsymbol{v}}_i, \overline{\boldsymbol{t}}_i) / \tau \,)}{\sum_{j=1}^{N} exp((\overline{\boldsymbol{v}}_i, \overline{\boldsymbol{t}}_j) / \tau \,)} \tag{4}$$

In the context of finer-grained alignment, comparative learning is still employed to maximize the similarity between correctly paired nouns (pairs of nouns) and, conversely, minimize the similarity between nouns that cannot be correctly paired. The model aims to maximize the similarity between x_n and y_n while minimizing the similarity between x_n and y_n. Here, x_n represents the representation of nouns captured from the video space, y_n represents the representation of correctly extracted nouns from the text, and y_k represents the representation of the sequence of other nouns extracted from the same batch of text.

This approach trains the multimodal encoder by relying on the video sequence representations to identify correctly paired nouns, compelling the video encoder to precisely capture the spatial content. The representation of the loss function is shown in Eq. (5):

$$L_2 = -\sum_{i=1}^{N} log \frac{exp(\, s(x_n^i, y_n^i) / \tau \,)}{\sum_{j=1}^{N} exp((x_n^i, y_k^j) / \tau \,)} \tag{5}$$

where, x_n^i and y_n^i denote the ith paired sample (positive sample), y_k^j denotes the negative sample of the ith sample, and N denotes the batch size.

Similarly, comparative learning for paired verbs focuses on maximizing the similarity between the verb representation x_v in the video space and the verb representation y_v in the text space. Simultaneously, it aims to minimize the similarity between x_v and other verb representations y_v in the text space. The loss function is expressed in Eq. (6):

$$L_3 = -\sum_{i=1}^{N} log \frac{exp(\, s(x_v^i, y_v^i) / \tau \,)}{\sum_{i=1}^{N} exp(\, s(x_v^i, y_p^j) / \tau \,)} \tag{6}$$

where, x_v^i and y_v^i denote the ith paired sample and y_p^i denotes the jth negative sample of the ith sample.

E. Objective function

This model finds the optimal model parameters by minimizing the sum of the three losses in Section III. *D.* The objective function is as in Eq. (7).

$$L = L_1 + L_2 + L_3 \tag{7}$$

F. Pre-training dataset

The video datasets employed for training this model are MSR-VTT [22] and MSVD [23], both comprising approximately 10K video data. To enhance the model's generalization, pre-training is conducted on the combined dataset of CC-3M [24] and WebVid-2M [1], resulting in approximately 5.5M video-text pairs after the merger.

IV. EXPERIMENTAL SETUP

In this study, our experiments aim to investigate whether the dual video-text alignment model enhances video retrieval accuracy. Specifically, the model is anticipated to achieve improved accuracy by incorporating a more nuanced understanding of both the video and description text, in contrast to the prevalent utilization of global features. Ablation experiments are then conducted to discern whether the enhanced retrieval accuracy is attributed to the finer-grained comprehension of the video and description text.

In this section, we initially present the general information and implementation details of the dataset used in the experiment. The model's performance is evaluated by reporting R@K and MdR, and the efficacy of our method is established through comparisons with other video retrieval approaches. We also detail the process and results of the ablation experiments. Additionally, this paper utilizes Grad-CAM [25] for generating class activation maps showcasing model cross-modal attention, and concludes with a case study illustrating the retrieval results of the model.

A. Datasets and Evaluation Metrics

1) Datasets: To benchmark against advanced baseline models and assess the performance of our proposed model, we conducted experiments on two widely used public datasets. The details of the dataset sources and divisions are outlined below:

MSR-VTT is curated with 257 popular queries from a commercial video search engine, encompassing a diverse array of categories and video content. It comprises 10k video clips and 200k descriptions. In previous work [26], the training set consists of 9k clip-text pairs, with the remaining 1k pairs designated for evaluation. This model follows the same division for training and evaluation.

MSVD is selected from YouTube, where each video description is independent and not influenced by the vocabulary or word order choices in previous descriptions. The dataset comprises 1,970 videos, ranging in length from 1 to 62 seconds. Each video is associated with approximately 40 descriptions. The training, validation, and test sets consist of 1200, 100, and 670 videos, respectively. This model undergoes training and evaluation using this standardized partition.

2) Evaluation metrics: To assess the performance of the model proposed in this paper, we employ standard evaluation metrics for video retrieval tasks: K recall (R@K, with K values of 1, 5, and 10, higher being preferable) and median rank (MdR, lower being preferable). R@K calculates the percentage of test samples with correct results within the top-K retrieval points relative to the query samples. Calculated as in Eq. (8):

$$R@K = \frac{TP}{TP + FN} \tag{8}$$

where, TP (True Positives) denotes the number of relevant videos that were correctly retrieved in the first K retrieval results and FN (False Negatives) denotes the number of relevant videos that were not retrieved in the first K results.

MdR measures the median position of the correct option in the sequence, assessing the model's capability to rank relevant videos effectively in the retrieval task.

B. Experimental Details

To facilitate training, the video size is initially adjusted to serve as the original input. Frames are sampled from a video during training, where the size of each patch is set to 16×16. Consequently, each video frame corresponds to one patch with sequence dimensions. The temporal and spatial attention blocks in the TimeSformer are initialized using ViT [8] weights pretrained on ImageNet-21k. The text encoder employs the Transformer architecture with eight attention heads, and the dimension of the common feature space is set to 256. During the training phase, this model utilizes the AdamW [28] optimizer with a learning rate set to 3×10^{-5} and 10 training epochs. Multi-interval learning rate tuning is applied: [4, 8], and weights are decayed to 0.1 times their original values.

Building upon previous research, pre-training using image-text pairs proves effective in enhancing the model's representation of the video space. The images in CC-3M were replicated and transformed into static videos. Additionally, we opted for the WebVid-2M video dataset, featuring 2.5M videos, for joint pre-training alongside CC-3M. This was accomplished using the AdamW optimizer, where the learning rate was set to 1×10^{-4}, the number of epochs was 20, and multi-interval learning rate tuning was applied [12, 16]. The weights were attenuated by a factor of 0.1 times their original values.

V. RESULTS AND DISCUSSION

To assess the impact of the video-text dual alignment model proposed in this paper on video retrieval accuracy, Tables I and II in this section present experimental results comparing this method with others on the MSR-VTT and MSVD datasets, with optimal results highlighted in bold. Through a comparison of evaluation metrics such as R@K and MdR, our method exhibits improvements in video retrieval task metrics over comparative models like X-CLIP and DCR.

A. Experimental Results

As depicted in Tables I and II, for the MSR-VTT and MSVD datasets, our method achieves a 1.3 percentage point increase in R@1 compared to previous state-of-the-art approaches. Notably, on the MSVD dataset, the improvement in R@1 is 2 percentage points. Simultaneously, there is a reduction in the MdR value in this task. This demonstrates that the incorporation of finer-grained alignment positively influences retrieval performance, underscoring the effectiveness of the proposed method.

The proposed method employs fine-grained alignment of words in text descriptions with actions or scenes in the video, leading to a more accurate alignment between video and text and improved modeling compared to the X-CLIP model, which outperformed other comparison models. While the ALPRO model introduces the concept of PEM for learning

fine-grained region-entity alignment, the fine-grained alignment in our method is notably more pronounced for the retrieval task. Moreover, the Clover model also incorporates the idea of modal alignment to enhance cross-modal feature alignment and fusion. In contrast, our approach utilizes the simpler dual alignment to achieve superior performance while successfully meeting the objective of improving retrieval accuracy outlined in this paper.

TABLE I. COMPARISON RESULTS WITH MAINSTREAM METHODS ON MSR-VTT DATASET

Methods	R@1/%	R@5/%	R@10/%	MdR
Frozen[1]	31.0	59.5	70.5	3.0
ALPRO[27]	33.9	60.7	73.2	3.0
Clover[29]	40.5	69.8	79.4	2.0
MELTR[33]	41.3	73.5	82.5	-
CLIP4Clip[30]	44.5	71.4	81.6	2.0
X-Pool[31]	46.9	72.8	82.2	2.0
X-CLIP[32]	49.3	75.8	84.8	2.0
TEFAL[36]	49.9	76.2	84.4	2.0
DCR[34]	50.2	76.6	84.7	**1.0**
Ours	**51.5**	**78.6**	**86.3**	2.0

TABLE II. COMPARISON RESULTS WITH MAINSTREAM METHODS ON MSVD DATASET

Methods	R@1/%	R@5/%	R@10/%	MdR
SupportSet[35]	28.4	60.0	72.9	4.0
Frozen[1]	33.7	64.7	76.3	3.0
CLIP4Clip[30]	46.2	76.1	84.6	2.0
DiffusionRet[37]	46.6	75.9	84.1	2.0
DMAE[38]	46.9	76.8	85.6	2.0
X-Pool[31]	47.2	77.4	86.0	2.0
DCR[34]	50.0	81.5	89.5	2.0
X-CLIP[32]	50.4	80.6	-	-
Ours	**52.4**	**83.3**	**90.5**	**1.0**

B. Ablation Study

To assess the efficacy of the dual alignment module and evaluate the impact of various comparison modules on retrieval outcomes, ablation experiments were conducted on two datasets, MSR-VTT and MSVD. Results from the ablation experiments are presented in Tables III and IV, while Fig. 3 provides a visual depiction for a more intuitive understanding of the effects of different alignment modules.

Initially, when solely engaged in the global video-text alignment task L_1, the efficiency of video retrieval across all combinations remains relatively low. This suggests a noticeable semantic gap between the global features of video and text modalities. Subsequently, with the inclusion of fine-grained noun alignment or verb alignment tasks ($L_1 + L_2$ or $L_1 + L_3$), the R@1 values on the MSR-VTT dataset improved by 2.8 and 4.2 percentage points, respectively. This indicates that the combination of global alignment and fine-grained alignment contributes to performance enhancement, albeit not significantly.

TABLE III. ABLATION EXPERIMENT ON MSR-VTT UNIT：%

Method	R@1	R@5	R@10
L_1	40.6	67.4	80.5
L_1+L_2	43.4	72.5	84.9
L_1+L_3	44.8	73.1	83.4
L_1+L_2+L_3	**51.5**	**78.6**	**86.3**

TABLE IV. ABLATION EXPERIMENT ON MSVD UNIT：%

Method	R@1	R@5	R@10
L_1	43.6	69.2	83.6
L_1+L_2	46.5	74.8	85.3
L_1+L_3	45.2	73.5	84.9
L_1+L_2+L_3	**52.4**	**83.3**	**90.5**

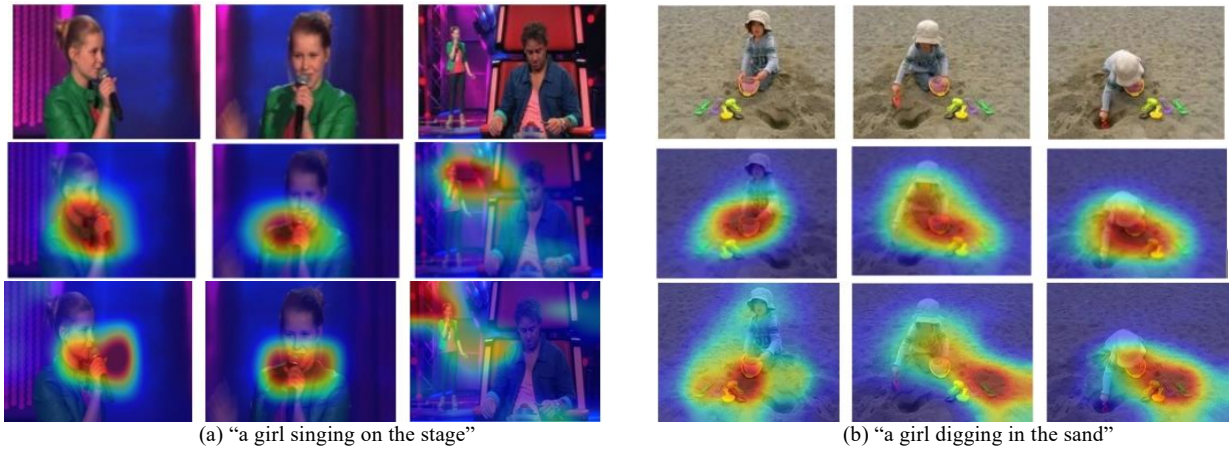
Ultimately, when integrating all three alignment tasks $L_1 + L_2 + L_3$, the model achieves optimal results on both datasets, as depicted in Fig. 3. The R@K values consistently rank highest when utilizing the three alignment strategies, reaching 51.5% and 52.4% at R@1 for the respective datasets. This underscores the effectiveness of combining three alignment strategies in improving retrieval efficiency. Attention Visualization

In the section on fine-grained video-text alignment, this model achieves a comprehensive understanding of the video content. Utilizing Grad-CAM, we generate class activation maps on the MSR-VTT dataset to visually represent the cross-modal attention between the video description and the video. This visualization aids in pinpointing corresponding regions in the video for the nouns or verbs mentioned in the text, showcasing the model's proficiency in fine-grained video-text alignment.

In this study, we employ Grad-CAM to visualize the third layer of the multimodal encoder. To enhance the presentation's clarity, we extract three frames from the video to illustrate the cross-modal attention between the verbs in the text and the video actions. We compare these results with the visualization outcomes of the X-CLIP model. In Fig. 3(a), depicting a scene where a girl sings on stage, our model's visualization captures the continuous focus on the girl's hand and face area, indicating alignment with the word "sing" in the text. Conversely, the X-CLIP model deviates from the girl's hand and face actions, favoring objects behind the girl. In Fig. 3(b), our model's visualization emphasizes the girl's hand movements corresponding to the word "digging" in the text, while the X-CLIP model seems more focused on objects beside the girl than her movements. This suggests that our model exhibits fine-grained cross-modal alignment capabilities compared to other models, emphasizing the importance of such alignment for improving retrieval results.

C. Video Retrieval Case

To demonstrate the retrieval effectiveness of the model, we visualize three examples of text retrieval on the test set of the MSR-VTT dataset, as shown in Fig. 4. In example (a), when searching for "two teams playing football," the model successfully retrieves scenes of people playing football, meeting the retrieval criteria. However, only one result is highlighted with a green box, as the model assumes one query corresponds to the optimal result. In Fig. 4(b), searching for "kids are singing by a table" yields the correct result in the first position. While other retrieval results are similar to the optimal one, they do not align with the scene described in the query. This highlights the model's ability to achieve fine-grained alignment between actions and scenes in the retrieval queries and video content, consequently enhancing retrieval efficiency.



(a) "a girl singing on the stage" (b) "a girl digging in the sand"

Fig. 3. Grad-CAM visualizes multimodal encoder cross-modal.

(a) Query: two teams playing football (b) Query: kids are singing by a table

Fig. 4. Text-video retrieval results example.

VI. Conclusion

This paper introduces an efficient video retrieval model through video-text alignment. The model uses TimeSformer and DistilBERT to extract unimodal feature representations from video and text, and performs global video-text alignment by linear projection and contrast learning. Subsequently, the local video information is compared and learned from the textual content by masking part of the textual information in order to achieve fine-grained video-text alignment. By enhancing the cross-modal training process and combining global and fine-grained alignment tasks, the model strengthens semantic associations between modal information, leading to improved alignment and enhanced video retrieval recall. Experiments on MSR-VTT and MSVD datasets validate the model's superiority and method effectiveness.

However, this method also has the non-negligible limitation that it takes a lot of time to perform video-text alignment, and it is also important to find a more efficient alignment.

In future work, we aim to delve deeper into exploring and integrating various modalities in videos, such as audio and subtitles, to further narrow the semantic gap between video and text and enhance the accuracy of video retrieval. Additionally, for the task of video retrieval, there is potential to train models tailored for retrieving videos in specific domains, making the models more specialized and efficient.

Acknowledgment

Henan Provincial Science and Technology Plan Project (No: 212102210417)

References

[1] M. Bain, A. Nagrani, G. Varol, & A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1728-1738, 2021.

[2] H. Luo, L. Ji, M. Zhong,, Y. Chen, W. Lei, N. Duan, & T, Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," Neurocomputing, 508, 293-304, 2022.

[3] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, & J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7331-7341, 2021.

[4] M. Wray, D. Larlus, G. Csurka, & D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," In Proceedings

of the IEEE/CVF international conference on computer vision, pp. 450-459, 2019.

[5] S. Chen, Y. Zhao, Q. Jin, & Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10638-10647, 2020.

[6] P. Wu, X. He, M. Tang, Y. Lv, & J. Liu, "Hanet: Hierarchical alignment networks for video-text retrieval," In Proceedings of the 29th ACM international conference on Multimedia , pp. 3518-3527, 2021, October.

[7] D. Kaufman, G. Levi, T. Hassner, & L. Wolf, "Temporal tessellation: A unified approach for video analysis," In Proceedings of the IEEE International Conference on Computer Vision, pp. 94-104, 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," In International Conference on Learning Representations. 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need.," Advances in neural information processing systems, 30, 2017.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... & I. Sutskever, "Learning transferable visual models from natural language supervision," In International conference on machine learning, pp. 8748-8763, PMLR, 2021, July.

[11] X. Li, F. Zhou, C. Xu, J. Ji,, & G. Yang, "Sea: Sentence encoder assembly for video retrieval by textual queries," IEEE Transactions on Multimedia, 23, pp. 4351-4362, 2020.

[12] C. Zhu, Q. Jia, W. Chen, Y. Guo, & Y. Liu, "Deep learning for video-text retrieval: a review," International Journal of Multimedia Information Retrieval, vol. 12, no 1, 2023.

[13] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, & J. Gao, "Unified vision-language pre-training for image captioning and vqa," In Proceedings of the AAAI conference on artificial intelligence, Vol. 34, No. 07, pp. 13041-13049, 2020, April.

[14] Y. Wang, & P. Shi, "Video-Text Retrieval by Supervised Multi-Space Multi-Grained Alignment," In Finding of the Association for Computational Linguistics: EMNLP, pp. 633-649, 2023.

[15] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, & X. Wang, "Dual encoding for zero-example video retrieval," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9346-9355, 2019.

[16] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, & T. S. Chua, "Tree-augmented cross-modal encoding for complex-query video retrieval," In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp. 1339-1348, 2020, July.

[17] Z. Wang, Y. Zhong, Y. Miao, L. Ma, & L. Specia, "Contrastive video-language learning with fine-grained frame sampling," In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, pp. 694-705, 2022.

[18] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong,, & S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," Advances in neural information processing systems, 34, 9694-9705, 2021.

[19] G. Bertasius, H. Wang, & L. Torresani, "Is space-time attention all you need for video understanding?". In ICML, Vol. 2, No. 3, p. 4, 2021, July.

[20] V. Sanh, L. Debut, J. Chaumond, & T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[21] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, 2019.

[22] J. Xu, T. Mei, T. Yao, & Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5288-5296, 2016.

[23] D. Chen, & W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp. 190-200, 2011, June.

[24] P. Sharma, N. Ding, S. Goodman, & R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556-2565, 2018, July.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, & D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," In Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.

[26] V. Gabeur, C. Sun, K. Alahari, & C. Schmid, "Multi-modal transformer for video retrieval.," In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 214-229, Springer International Publishing, 2020.

[27] D. Li, J. Li, H. Li, J. C. Niebles, & S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4953-4963, 2022.

[28] I. Loshchilov, F. Hutter, "Decoupled weight decay regularization," International Conference on Learning Representations, 2017.

[29] J. Huang, Y. Li, J. Feng, X. Wu, X. Sun, & R. Ji, "Clover: Towards a unified video-language alignment and fusion model," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14856-14866, 2023.

[30] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, & T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," Neurocomputing, pp. 293-304, 2022.

[31] S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, & G. Yu, "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, 2022.

[32] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, & R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," In Proceedings of the 30th ACM International Conference on Multimedia, pp. 638-647, 2022, October.

[33] D. Ko, J. Choi, H. K. Choi, K. W. On, B. Roh, & H. J. Kim, "MELTR: Meta Loss Transformer for Learning to Fine-tune Video Foundation Models," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20105-20115, 2023.

[34] Q. Wang, Y. Zhang, Y. Zheng, P. Pan, & X. S. Hua, "Disentangled representation learning for text-video retrieval," arXiv preprint arXiv:2203.07111, 2022.

[35] M. Patrick, P. Y. Huang, Y. Asano, F. Metze, A. Hauptmann, J. Henriques, & A. Vedaldi, "Support-set bottlenecks for video-text representation learning," International Conference on Learning Representations, 2021

[36] S. Ibrahimi, X. Sun, P. Wang, A. Garg, A. Sanan, & M. Omar, "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, 2023.

[37] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, ... & J. Chen, "Diffusionret: Generative text-video retrieval with diffusion model," International Conference on Computer Vision, pp. 2470-2481, 2023.

[38] C. Jiang, H. Liu, X. Yu, Q. Wang, Y. Cheng, J. Xu, ... & Y. Qi, "Dual-Modal Attention-Enhanced Text-Video Retrieval with Triplet Partial Margin Contrastive Learning," In Proceedings of the 31st ACM International Conference on Multimedia, pp. 4626-4636, 2023.