Towards Efficient and Effective Text-to-Video Retrieval with Coarse-to-Fine Visual Representation Learning

Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, Han Li

Kuaishou Technology
{tiankaibin, chengyanhua, liuyi24, houxinglin, chenquan06, lihan08}@kuaishou.com

Abstract

In recent years, text-to-video retrieval methods based on CLIP have experienced rapid development. The primary direction of evolution is to exploit the much wider gamut of visual and textual cues to achieve alignment. Concretely, those methods with impressive performance often design a heavy fusion block for sentence (words)-video (frames) interaction, regardless of the prohibitive computation complexity. Nevertheless, these approaches are not optimal in terms of feature utilization and retrieval efficiency. To address this issue, we adopt multi-granularity visual feature learning, ensuring the model's comprehensiveness in capturing visual content features spanning from abstract to detailed levels during the training phase. To better leverage the multi-granularity features, we devise a two-stage retrieval architecture in the retrieval phase. This solution ingeniously balances the coarse and fine granularity of retrieval content. Moreover, it also strikes a harmonious equilibrium between retrieval effectiveness and efficiency. Specifically, in training phase, we design a parameter-free text-gated interaction block (TIB) for fine-grained video representation learning and embed an extra Pearson Constraint to optimize cross-modal representation learning. In retrieval phase, we use coarse-grained video representations for fast recall of top-k candidates, which are then reranked by fine-grained video representations. Extensive experiments on four benchmarks demonstrate the efficiency and effectiveness. Notably, our method achieves comparable performance with the current state-of-the-art methods while being nearly 50 times faster.

Introduction

With the explosive growth of videos uploaded online every day from platforms like TikTok, Kwai, YouTube, and Netflix, text-to-video retrieval is a crucial and fundamental task for multi-modal representation learning (Fang et al. 2021; Ge et al. 2022; Gorti et al. 2022; Luo et al. 2022; Ma et al. 2022). Recently, the pre-trained text-image matching models (CLIP (Radford et al. 2021)) from a large scale of web-collected image-text pairs show the great success on diverse vision-language downstream tasks (Nichol et al. 2021; Ramesh et al. 2022; Mokady, Hertz, and Bermano 2021; Hu et al. 2022b; Li et al. 2022; Conde and Turgutlu 2021). In light of the well-learned visual features, a preliminary study is conducted by CLIP4Clip (Luo et al. 2022), which transfers the pre-trained CLIP to video-language domain with simple MeanPooling and achieves a certain level of promotion. A problem arises where, unlike relatively less visual information in text-image matching, aggregating the entire video as a representation in text-to-video retrieval can lead to over-abstraction and be misleading. As one sentence generally describes video sub-regions of interest. Therefore, a natural idea is to consider how to align text and video representations at a finer granularity. The development of CLIP variants can be divided into two main categories to address the above problem. One category focuses on designing a heavy fusion block to strengthen the interaction between visual (video, frames) and text (sentence, words) cues for better alignment of the two modalities (Liu et al. 2021; Ma et al. 2022). The other one optimizes text-driven video representations by keeping multi-grained features including video-level and frame-level for brute-force search (Gorti et al. 2022; Wang et al. 2022).

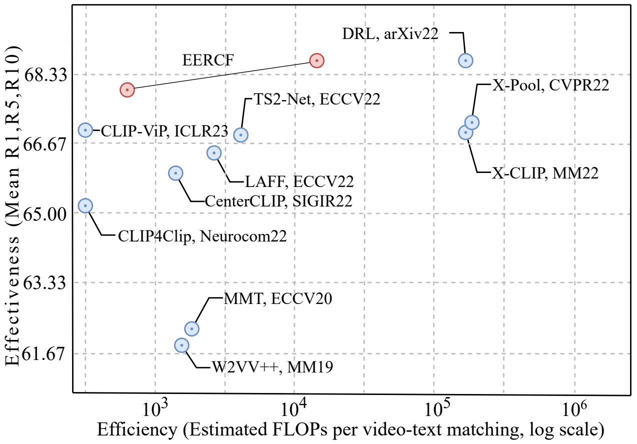

Figure 1: Effectiveness and efficiency for text-to-video retrieval models. We evaluate our approach under the settings of MSRVTT-1K-Test and backbone CLIP(ViT-B/32). The current trend of mainstream is reflected from the lower left to the upper right corner. Our method achieves the best balance, positioned at the upper left corner.

Despite the big promotion, such mechanisms bring a non-negligible increase in computational cost for real applications, as shown in Fig.1. The computational cost here refers to text-video similarity calculation, when a video (text) is represented by one or more vectors. This determines the efficiency of retrieval in practical applications. Additionally, we also believe that excessively fine-grained calculations may amplify the noise in local parts of the video, resulting in reduced retrieval effectiveness.

To make a better trade-off between effectiveness and efficiency for text-to-video retrieval, this paper proposes a novel method, namely EERCF, towards coarse-to-fine adaptive visual representations learning following a recall-then-rerank pipeline. Towards coarse-to-fine adaptive visual representations, we adopt the basic framework of CLIP4Clip (Luo et al. 2022) and design a parameter-free text-gated interaction block (TIB) for fine-grained video representation learning, which further refines the granularity from frame to patch compared with other methods. Concretely, TIB can adjust the weights of different frame-level features or patch-level features given a text and aggregate them into a video representation, making the matching between text and video more accurate. In the learning process, we employ a joint inter- and intra-feature supervision loss following DRL (Wang et al. 2022), satisfying both cross-modal feature matching and redundancy reduction across feature channels. Inspired by the success of Pearson Constraint in knowledge distillation (Huang et al. 2022), we adapt it to reduce redundancy across feature channels.

Taking efficiency into consideration, a two-stage retrieval strategy is adopted when using coarse-to-fine visual representations in practice. There are three levels of video representations, including text-agnostic features without text interaction and text-driven aggregation of frame-level and patch-level features generated from TIB module. The text-agnostic video representations are used for fast recall of top-k candidates, which are then reranked by another two fine-grained video representations. Besides, we found that using coarse-grained features in the recall stage can avoid the noise introduced by overly fine-grained features which typically pay more attention to visual details, thereby improving retrieval performance.

Therefore, our approach is towards **E**fficient and **E**ffective text-to-video **R**etrievl with **C**oarse-to-**F**ine visual representation learning and is coined **EERCF**. We summarize our main contributions as follows:

1) We introduce a text-gated interaction block without extra learning parameters for multi-grained adaptive representation learning, whilst introducing a combination of inter-feature contrastive loss and intra-feature Pearson Constraint for optimizing feature learning.

2) We propose a two-stage text-to-video retrieval strategy that strikes the optimal balance between effectiveness and efficiency, facilitating the practical implementation.

3) Our method EERCF achieves comparable performance with the current state-of-the-art(SOTA) methods while our FLOPs for cross-modal similarity matching in MSRVTT-1K-Test, MSRVTT-3K-Test, VATEX, and ActivityNet are 14, 39, 20 and 126 times less than the SOTA.

Related Work

Cross-model Representation Optimization

The pioneering work, CLIP (Radford et al. 2021), collects 400M public image-text pairs on the internet and demonstrates the great power of visual-linguistic representation on various downstream tasks, including text-to-image generation (Nichol et al. 2021; Ramesh et al. 2022), image caption (Mokady, Hertz, and Bermano 2021; Hu et al. 2022b) and vision understanding (Li et al. 2022; Conde and Turgutlu 2021). Benefiting from the pre-trained CLIP model, CLIP4Clip (Luo et al. 2022) adapts it to video-text retrieval with MeanPooling for aggregating video features while still outperforming models pre-trained on video data (Bain et al. 2021; Xu et al. 2021; Xue et al. 2022). CLIP and its variants all employ contrastive learning for cross-modal feature alignment. Recently, DRL (Wang et al. 2022) demonstrates the optimization of correlation reduction for video-text retrieval via a regularization trick. Therefore, We embed the extra Pearson Constraint as a trick to reduce correlation among different cross-modal feature channels, resulting in optimized video and text features.

Boosting Video Representation from Text Interaction

Note that some CLIP variants, such as CLIP4Clip, TS2-Net, and CLIP-VIP (Luo et al. 2022; Xue et al. 2023), for video-text retrieval utilize two independent encoders to represent the two modalities for efficient deployment. For lack of text interaction, the video representation can be over-abstract and misleading to match the common sub-cues depicted by multiple corresponding texts. To address this gap, recent works, such as X-CLIP, TS2-Net, X-Pool and DRL (Ma et al. 2022; Liu et al. 2022; Gorti et al. 2022; Wang et al. 2022), pay attention to adaptive video features with different text interaction mechanisms. X-CLIP and TS2-Net (Ma et al. 2022) design a heavy interaction block with multi-grained attention for joint video and text representation learning. X-Pool and DRL (Gorti et al. 2022; Wang et al. 2022) improve performance by redesigning the interaction block with a few learning parameters while computing similarity measures between multiple frame features and the query sentence feature for exhaustive search. Our approach goes one step further, developing a text-gated interaction block without extra learning parameters and generating more fine-grained levels of video features, including a text-agnostic version and text-driven aggregation of patch-level and frame-level versions. The introduction of more fine-grained features has a worse impact on efficiency. Therefore, it is inevitable to adopt an efficient two-stage strategy.

Re-ranking for Cross-Modal Retrieval

Re-ranking in cross-modal retrieval is not uncommon. For instance, in text-image retrieval tasks, (Wei et al. 2020) proposes in the embedding space MVSE++ to rerank the recalled candidates using k nearest neighbors. This is a typical unsupervised re-ranking method, as the re-ranking process is independent of the text modality. However, in the field of text-video retrieval, there are not many methods

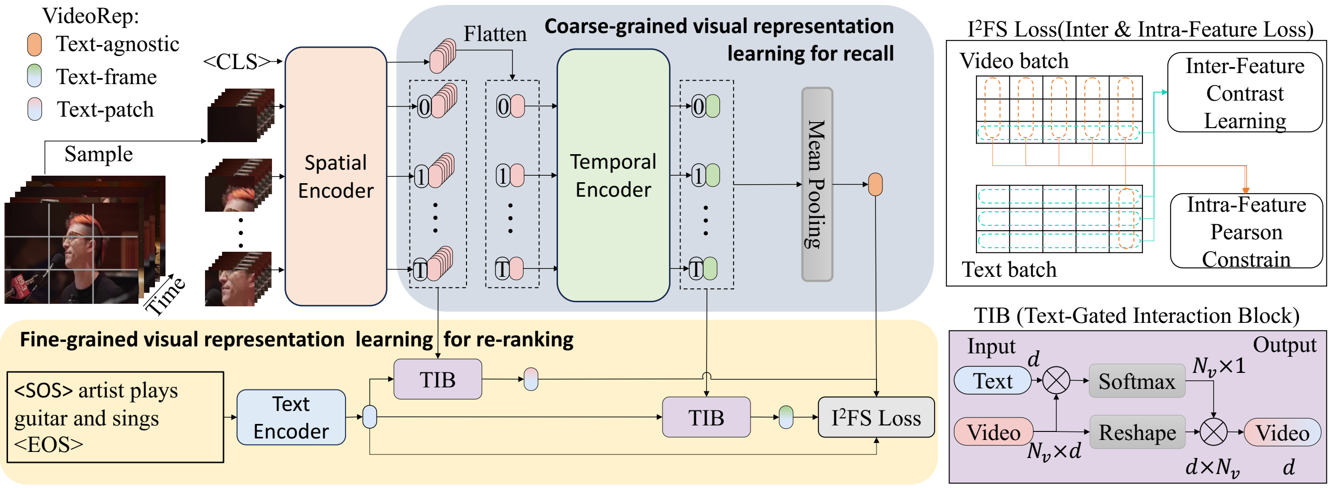
VideoRep:
Text-agnostic
Text-frame
Text-patch
<CLS>
Sample
Time
Spatial Encoder
Flatten
Coarse-grained visual representation learning for recall
Temporal Encoder
Mean Pooling
0
1
T
I²FS Loss(Inter & Intra-Feature Loss)
Video batch
Inter-Feature Contrast Learning
Intra-Feature Pearson Constrain
Text batch
Fine-grained visual representation learning for re-ranking
<SOS> artist plays guitar and sings <EOS>
Text Encoder
TIB
TIB
I²FS Loss
TIB (Text-Gated Interaction Block)
Input
Output
Text
d
Softmax
N_v×1
Video
Reshape
Video
N_v×d
d×N_v
d

Figure 2: Overview of the proposed EERCF framework. EERCF mainly consists of two parts: 1) Coarse-grained and fine-grained visual representations obtained from the TIB module for the recall-reranking pipeline. 2) Inter- and intra-feature supervision loss for optimizing representation learning. Best viewed in color.

for re-ranking. Our method provides a paradigm for implementing re-ranking. Specifically, the text-agnostic coarse-grained video features are used for fast recall, and the text-driven fine-grained features benefit to high-performance re-ranking, thus making a better trade-off between efficiency and effectiveness. As the re-ranking process is related to the text modality, we refer to it as text-supervised re-ranking.

Task Formulation

This paper mainly focuses on the task of text-to-video retrieval ($t2v$), while also taking into consideration the task of video-to-text retrieval ($v2t$). In $t2v$, the objective is to rank all videos from a video gallery set $\mathcal{V}$ given a query text t based on a similarity score function $s(t, v)$. The $v2t$ task is the reverse of $t2v$. In both tasks, the gallery set is provided ahead of time for retrieval. Specifically, a video v consists of T sequential frames $\{f_1, f_2, ..., f_T | f_i \in \mathbb{R}^{H\times W\times C}\}$, where each frame is divided into N patches $\{f_i^1, f_i^2, ..., f_i^N | f_i^n \in \mathbb{R}^{P\times P\times C}\}$ with $P \times P$ size. A text t is defined as a sequence of tokenized words.

Methodology

In this section, we first introduce the overall architecture of coarse-to-fine visual representation learning in "Overall Architecture". Then we provide a detailed exposition of the parameter-free text-gated interaction block(TIB). Next, we explain the joint inter- and intra-feature supervision loss function to optimize cross-modal feature learning. Finally, we present a two-stage retrieval strategy utilizing well-extracted multi-grained features.

Overall Architecture

Fig.2 illustrates the overall architecture of the proposed EERCF framework. There are three levels of video representations that need to be learned. The first one is a text-agnostic video representation. We follow the same setting of CLIP4Clip (Luo et al. 2022), which utilizes a spatial encoder (SE) with 12 transformer layers initialized by the public CLIP checkpoints, followed by a temporal encoder (TE) with 4 transformer layers to model temporal relationship among sequential frames and a MeanPooling layer (MP) to aggregate all frame-level features into a text-agnostic feature vector. The procedure above can be formulated as follows:

$$\varphi([f_i^0; f_i^1; ...; f_i^N]) = \mathbf{SE}([f_i^0; f_i^1; ...; f_i^N] + E_{spos}) \tag{1}$$

$$\phi([f_1; f_2; ...; f_T]) = \mathbf{TE}([\varphi(f_1^0); ...; \varphi(f_T^0)] + E_{tpos}) \tag{2}$$

$$\boldsymbol{v}_{L_1} = \mathbf{MP}(\phi(f_1), \phi(f_2), ..., \phi(f_T)) \tag{3}$$

where $\varphi(f_i^n) \in \mathbb{R}^D$ denotes the n_{th} patch feature of the i_{th} frame except that $\varphi(f_i^0)$ is the [CLS] token prediction to encode all patch features within i_{th} frame, $\phi(f_i) \in \mathbb{R}^D$ denotes the i_{th} frame feature, and $\boldsymbol{v}_{L_1} \in \mathbb{R}^D$ is the resulted text-agnostic video feature. We add spatial position embeddings $E_{spos} \in \mathbb{R}^D$ to all the patch embeddings. Then all the outputs $\varphi(f_i^0)$ along with temporal position embeddings $E_{tpos} \in \mathbb{R}^D$ are loaded into TE to learn temporal relations among frames of a video. To further capture particularly visual cues conditioned on the text, another two video features are extracted based on a text-gated interaction block (TIB):

$$\boldsymbol{v}_{L_2} = \mathbf{TIB}(\phi(f_1), \phi(f_2), ..., \phi(f_T), \theta(t)) \tag{4}$$

$$\boldsymbol{v}_{L_3} = \mathbf{TIB}(\varphi(f_1^1), ..., \varphi(f_1^N), ..., \varphi(f_T^1), ..., \varphi(f_T^N), \theta(t)) \tag{5}$$

where $\theta(t) \in \mathbb{R}^D$ denotes the sentence feature of the query text, which drives the aggregation of all the frame features to capture related spatiotemporal visual cues, resulting in text-frame interaction feature $\boldsymbol{v}_{L_2} \in \mathbb{R}^D$, as well as the aggregation of all the patch features to keep more fine-grained and aligned visual cues, building text-patch interaction feature $\boldsymbol{v}_{L_3} \in \mathbb{R}^D$. Next, we will provide a detailed introduction to learning finer-grained video representations using the TIB module.

Text-Gated Interaction Block

The TIB module is a parameter-free text-gated interaction block to align the fine-grained video features with the query sentence feature, as shown in Fig.2. It is a simple attention mechanism based on the Softmax function with a temperature coefficient. Then the text-driven video features can be rewritten as follows:

$$\boldsymbol{v}_{L_2} = \sum_{i=1}^{T} Softmax(\phi(f_i)^{\top}\theta(t)/\pi)\phi(f_i) \tag{6}$$

$$\boldsymbol{v}_{L_3} = \sum_{i=1}^{T}\sum_{m=1}^{N} Softmax(\varphi(f_i^m)^{\top}\theta(t)/\pi)\varphi(f_i^m) \tag{7}$$

where π is the temperature, which decides how much visual cues will be kept for video-level feature aggregation based on the softmax similarity from a text to all frames or patches. A small value of π only emphasizes those most relevant visual cues, while a large value pays attention to much more visual cues. Considering efficiency, we do not introduce any learning parameters in TIB.

Inter- and Intra-Feature Supervision Loss

We train EERCF using mini-batch iterations with each batch of B video-text pairs $\{(v_b, t_b)\}_{b=1}^{B}$. In each pair, the text t_b is a corresponding description of the video v_b. For each modality, we extract a feature matrix $\boldsymbol{F} \in \mathbb{R}^{B\times D}$, where a row vector $\boldsymbol{F}_{b,:}$ denotes the feature representation of instance b, and a column vector $\boldsymbol{F}_{:,d}$ denotes all instances' feature values at channel d. The two types of vectors are separately used for inter-feature alignment and intra-feature correlation reduction between video and text via the following loss functions.

Contrastive Loss for Inter-Feature Supervision. Contrastive learning is popularly applied in multi-modal learning tasks (Radford et al. 2021; Luo et al. 2022). Similarly, we employ the infoNCE loss by considering video-text matching pairs as positives and other non-matching pairs in the batch as negatives. Specifically, we jointly optimize the symmetric video-to-text and text-to-video losses:

$$\mathcal{L}_{inter}^{t2v} = \frac{1}{B}\sum_{b_1=1}^{B} InfoNCE(\boldsymbol{F}_{b_1,:}^{(v)}, \boldsymbol{F}^{(t)}) \tag{8}$$

$$\mathcal{L}_{inter}^{v2t} = \frac{1}{B}\sum_{b_1=1}^{B} InfoNCE(\boldsymbol{F}_{b_1,:}^{(t)}, \boldsymbol{F}^{(v)}) \tag{9}$$

$$\mathcal{L}_{inter} = \mathcal{L}_{inter}^{v2t} + \mathcal{L}_{inter}^{t2v} \tag{10}$$

Algorithm 1: Recall and Re-ranking during Retrieval

```
**Input**: video gallery set: $\mathcal{V} = \{v_i\}_{i=1}^{K}$, a query text: $t$
**Output**: the most matching video
Encode $t$ as a text feature $\theta(t)$
# Reacll Stage
for $i \leftarrow 1, K$ do
    Encode $v_i$ as a video-level video feature $\boldsymbol{v}_{L_1,i}$
    Compute the similarity score between $\theta(t)$ and $\boldsymbol{v}_{L_1,i}$
end for
Select top $k$ highest score videos as the candidate set.
# Re-ranking Stage
for $i \leftarrow 1, k$ do
    Encode $v_i$ as a frame-level video feature $\boldsymbol{v}_{L_2,i}$
    Encode $v_i$ as a patch-level video feature $\boldsymbol{v}_{L_3,i}$
    Compute the similarity score between $\theta(t)$ and
    weighted sum $(\boldsymbol{v}_{L_1,i}, \boldsymbol{v}_{L_2,i}, \boldsymbol{v}_{L_3,i})$
end for
Select the highest score video as the most matching
video
```

Pearson Constraint for Intra-Feature Supervision. Pearson Constraint is successfully exploited for knowledge distillation in research (Huang et al. 2022). In this paper, we transfer the idea of correlation reduction among intra-features as a trick to optimize feature learning. Concretely, Pearson Constraint is defined as a distance measure function:

$$d_p(\boldsymbol{F}_{:,d_1}^{(v)}, \boldsymbol{F}_{:,d_2}^{(t)}) = 1 - \rho_p(\boldsymbol{F}_{:,d_1}^{(v)}, \boldsymbol{F}_{:,d_2}^{(t)}) \tag{11}$$

where $\rho_p(\boldsymbol{F}_{:,d_1}^{(v)}, \boldsymbol{F}_{:,d_2}^{(t)})$ is the Pearson coefficient[1] among normalized intra-feature channels between video and text. Finally, we define the intra-feature loss function as follows:

$$\begin{aligned}\mathcal{L}_{intra} = &\sum_{d=1}^{D}\left\|d_p(\boldsymbol{F}_{:,d}^{(v)}, \boldsymbol{F}_{:,d}^{(t)})\right\|^2 \\ &+ \alpha\sum_{d_1=1}^{D}\sum_{d_2\neq d_1}\left\|1 - d_p(\boldsymbol{F}_{:,d_1}^{(v)}, \boldsymbol{F}_{:,d_2}^{(t)})\right\|^2\end{aligned} \tag{12}$$

where $\|...\|^2$ denotes L2-norm regularization and α controls the magnitude of correlation reduction term. $\mathcal{L}_{intra}$ can benefit the model by learning particular and orthogonal cues among each feature channel. Pearson Constraint achieves a relaxed correlation reduction by allowing each feature channel to have the strongest correlation with itself, without needing to be completely independent of other channels. It benefits the model by learning compact video features.

Total Loss Function. As a result, the overall training loss $\mathcal{L}_{all}$ can be composed of the inter-feature and intra-feature supervision losses, $i.e.$,

$$\mathcal{L}_{all} = \sum_{\boldsymbol{v}\in\{\boldsymbol{v}_{L_1}, \boldsymbol{v}_{L_2}, \boldsymbol{v}_{L_3}\}} \lambda_{\boldsymbol{v}}(\mathcal{L}_{inter} + \beta\mathcal{L}_{intra}) \tag{13}$$

where β weights the loss importance between $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$, and $\lambda_{\boldsymbol{v}}$ balances the contribution of each level of video feature learning.

Two-stage Strategy in Retrieval

To balance the efficiency and effectiveness for text-to-video retrieval, we utilize $\boldsymbol{v}_{L_1}$ for fast recall of top-k condidates

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Model	$t2v$ Retrieval				$v2t$ Retrieval				FLOPs
	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	(k=1000)
Backbone model: ViT-B/32									
TeachText (Croitoru et al. 2021)	29.6	61.6	74.2	55.1	32.1	62.7	75.0	56.6	0.8k
SEA (Li et al. 2020)	37.2	67.1	78.3	60.9	-	-	-	-	10.2k
W2VV++ (Li et al. 2019)	39.4	68.1	78.1	61.9	-	-	-	-	2.0k
BridgeFormer (Ge et al. 2022)	44.9	71.9	80.3	65.7	-	-	-	-	0.5k
LAFF (Hu et al. 2022a)	45.8	71.5	82.0	66.4	-	-	-	-	4.1k
CLIP4Clip†	42.8	71.6	81.1	65.2	41.4	70.6	80.5	64.2	**0.5k**
CenterCLIP (Zhao et al. 2022)	44.0	70.7	81.4	65.4	42.9	71.4	81.7	65.3	1.5k
X-CLIP†	46.3	72.1	81.8	66.7	**45.9**	72.8	81.2	66.7	220.9k
TS2-Net†	46.7	72.6	81.2	66.8	43.6	71.1	82.7	65.8	6.1k
CLIP-VIP (Xue et al. 2023)	46.5	72.1	82.5	67.0	40.6	70.4	79.3	63.4	**0.5k**
X-Pool (Gorti et al. 2022)	46.9	72.8	82.2	67.3	44.4	73.3	**84.0**	67.2	275.0k
DRL (Wang et al. 2022)	47.4	**74.6**	83.8	**68.6**	45.3	73.9	83.3	67.5	220.4k
EERCF (ours)	**47.8**	74.1	**84.1**	**68.6**	44.7	**74.2**	83.9	**67.6**	16.0k
Backbone model: ViT-B/16									
CLIP4Clip†	46.4	72.1	82.0	66.8	45.4	73.4	82.4	67.1	**0.5k**
X-CLIP†	49.3	75.8	84.8	70.0	48.9	76.8	84.5	70.1	220.9k
DRL (Wang et al. 2022)	50.2	76.5	84.7	70.5	48.9	76.3	85.4	70.2	220.4k
DRL* (Wang et al. 2022)	53.3	**80.3**	**87.6**	**73.7**	**56.2**	**79.9**	**87.4**	**74.5**	220.4k
EERCF (ours)	49.9	76.5	84.2	70.2	47.8	75.3	84.2	69.1	0.8k
EERCF* (ours)	**54.1**	78.8	86.9	73.2	55.0	77.8	85.7	72.8	0.8k

Table 1: Comparison of retrieval efficiency and effectiveness on the MSRVTT-1K-Test. The best results are shown in bold and the results unavailable are left blank. Methods marked with † are reproduced in this paper with the same experimental settings for fair comparison. * denotes we add the DSL or Q-Norm trick to achieve the best performance in comparison.

and then rerank them via $\boldsymbol{v}_{L_2}$ and $\boldsymbol{v}_{L_3}$. Note that all video and text features are L_2 normalized for similarity measure. It is worth mentioning that we consider using a two-stage method, partly due to its efficiency improvement, and partly due to the fact that overly fine-grained features may lead to an excessive focus on local noise. The two-stage approach strikes a good balance between overly abstract and overly detailed video representations.

Experiments

We perform experiments on the commonly used benchmark of MSR-VTT (Xu et al. 2016), VATEX (Wang et al. 2019), MSVD (Chen and Dolan 2011), and ActivityNet (Heilbron et al. 2015). These datasets vary in video duration, content, and text annotations, providing a comprehensive evaluation of different methods. Details are listed in the supplementary material. Following existing literature (Croitoru et al. 2021; Li et al. 2020, 2019; Ge et al. 2022; Hu et al. 2022a; Luo et al. 2022; Zhao et al. 2022; Gorti et al. 2022; Ma et al. 2022; Xue et al. 2023; Liu et al. 2022), we report Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), and mean result of them (Mean) for comparison. We use FLOPs to evaluate efficiency for text-video similarity calculation, which is calculated by THOP[2]. It should be noted that due to the relatively small size of the MSVD(Chen and Dolan 2011), the results are in the supplementary material.

Implementation Details

We perform the experiments on 24 NVIDIA Tesla T4 15GB GPUs using the PyTorch library. Similar to (Luo et al. 2022; Ma et al. 2022), the spatial encoder and text encoder of EERCF are initialized by the CLIP checkpoints. We train our model via Adam optimizer and decay the learning rate using a cosine schedule strategy. For better finetuning, we set different learning rates for different modules, where the spatial encoder and text encoder are set to 1e-7, owning to CLIP initialization, and other new modules, like the temporal encoder, are set to 1e-4. The max word token length and max frame length are fixed to 32 and 12 for MSR-VTT, MSVD, and VATEX, while the corresponding settings are 64 and 64 for ActivityNet due to longer captions of video-paragraph retrieval. Limited by GPU memory, we set the batch size of MSR-VTT, MSVD, VATEX, and ActivityNet to 240, 240, 360, and 96, respectively. We train 5 epochs for all datasets. Unless otherwise specified, the hyperparameters mentioned in our equations are empirically set as follows: $\pi = 0.1$ and $\pi = 0.01$ separately for frame-level and patch-level TIB module, $\{\alpha = 0.05\}$ in $\mathcal{L}_{intra}$ loss, $\{\beta = 0.001, \lambda_{\boldsymbol{v}_{L_1}} : \lambda_{\boldsymbol{v}_{L_2}} : \lambda_{\boldsymbol{v}_{L_3}} = 5 : 5 : 1\}$ in the total loss, and top-k=50 for our coarse-to-fine retrieval.

Performance Comparison

On all the datasets, EERCF can achieve a performance close to or even exceed the SOTA methods while maintaining the advantage of retrieval efficiency. We present a detailed com-

[2] https://github.com/Lyken17/pytorch-OpCounter

Model	MSRVTT-3K-Test				VATEX				ActivityNet			
	R@1	R@5	Mean	FLOPs	R@1	R@5	Mean	FLOPs	R@1	R@5	Mean	FLOPs
TeachText	15.0	38.5	35.1	0.8k	53.2	87.4	78.0	0.8k	23.5	57.2	58.9	0.8k
SEA	19.9	44.3	40.2	10.2k	52.4	90.2	79.5	10.2k	-	-	-	-
W2VV++	23.0	49.0	44.2	2.0k	55.8	91.2	81.0	2.0k	-	-	-	-
LAFF	29.1	54.9	49.9	4.1k	59.1	**91.7**	82.4	4.1k	-	-	-	-
CLIP4Clip†	29.4	54.9	50.0	**0.5k**	61.6	91.1	82.8	**0.5k**	39.7	71.0	64.7	**0.5k**
TS2-Net†	29.9	56.4	51.2	6.1k	61.1	91.5	82.9	6.1k	37.3	69.9	63.5	32.8k
X-CLIP†	31.2	**57.4**	**52.2**	220.9k	62.1	90.8	82.7	220.9k	**44.4**	**74.6**	68.0	2175.9k
DRL	-	-	-	-	**63.5**	**91.7**	**83.9**	220.4k	44.2	74.5	**68.3**	2175.4k
EERCF (ours)	**31.5**	**57.4**	**52.2**	5.7k	62.6	91.5	83.3	10.8k	43.1	74.5	67.9	17.3k

Table 2: Comparison of text-to-video retrieval efficiency and effectiveness on the MSRVTT-3K-Test, VATEX and ActivityNet. Results on video-to-text retrieval are similar and omitted due to limited space.

Model	R@1	R@5	R@10	Mean
EERCF	**47.8**	**74.1**	**84.1**	**67.8**
- w/o patch-level feature	47.2	73.2	82.2	67.5
- w/o frame-level feature	45.0	71.4	81.4	65.9
- w/o all text-driven features	42.8	71.6	81.1	65.2

Table 3: Ablation for fine-grained video features in EERCF on MSRVTT-1K-Test dataset.

parison of both efficiency and effectiveness on MSRVTT-1K-Test in Tab.1. We conclude the following observation:

- Building on ViT-B/32, EERCF achieves outstanding results of 68.6/67.6 Mean in the $t2v/v2t$ task, surpassing stronger competitors such as CLIP-VIP (the baseline version without additional data), X-Pool, X-CLIP, TS2Net and DRL. EERCF achieves a large gain than CLIP4Clip model by +11.7% (+5.0%) relative (absolute) improvement on $t2v$ R@1, and +8.0% (+3.3%) improvement on $v2t$ R@1.
- Building on ViT-B/16, remarkably, EERCF achieves performance comparable to that of DRL while having a significantly lower computation cost of only 0.8k FLOPs. Specifically, the computation cost of DRL is approximately 275 times higher than that of EERCF.

To further validate the generalization of EERCF, we evaluate its performance on MSRVTT-3K-Test, VATEX and ActivityNet, as presented in Tab.2. EERCF achieves considerable performance improvement in an efficient manner, being nearly 39, 20, and 126 times faster than the SOTAs in MSRVTT-3K-Test, VATEX, and ActivityNet.

Complexity Analysis

Normally, we denote the computational complexity as $\mathcal{O}(ND)$ for dot product retrieval, where both video and text are represented by a D dimension vector and the gallery set is with size N. When performing more fine-grained retrieval, we set the number of frames as N_v, the number of words as N_t, the number of patches per frame as N_p, and the number of candidate set as N_r, which is much smaller than N. Our complexity, $\mathcal{O}(ND + N_r(1 + N_v + N_p)D)$, is significantly better than DRL's $\mathcal{O}(NN_vN_tD)$ and X-CLIP's $\mathcal{O}(N(N_vN_t + N_v + N_t + 1)D)$.

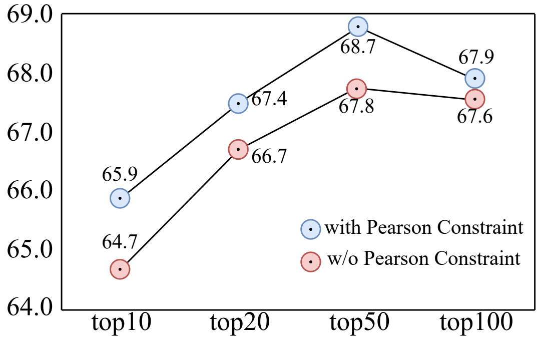

Figure 3: Retrieval performance on MSRVTT-1K-Test based on different number of re-ranking candidates k.

Ablation Study

Since MSR-VTT is more popular and competitive compared to other datasets, we conduct ablation, quantitative and qualitative experiments on it. In this section, we carefully investigate the proposed EERCF, including the contributions of multi-grained video representations, the effect of Pearson Constraint, the selection of top-k re-ranking hyperparameter. We further compare the differences between different re-ranking methods and visualize how fine-grained visual representations affect the re-ranking process.

Fine-grained text-driven features. Ablation results are presented in Tab.3. Upon removing either the text-frame interaction feature or the text-patch interaction feature from EERCF, a decline in retrieval performance can be observed, particularly in the removal of the text-frame feature. This decline serves to underscore the effectiveness of both fine-grained features, which adeptly capture distinct levels of visual and temporal cues from the query text. Furthermore, when we exclude all text-driven video features in EERCF and instead opt for the conventional single-stage retrieval paradigm relying solely on the text-agnostic video feature, a significant decrease in retrieval performance becomes evident. The phenomenon demonstrates the limitations inherent in the text-agnostic video feature, which may tend to oversimplify and potentially mislead the matching process for

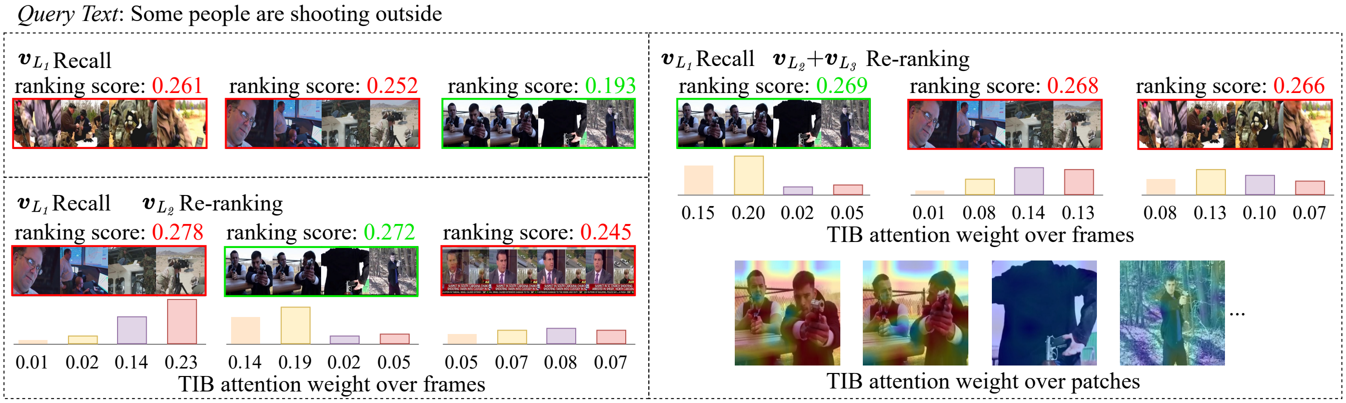

Figure 4: Visualization of the coarse-to-fine retrieval process on MSRVTT-1K-Test. Green boxes mean the ground truth video corresponding to the query text, and red boxes denote confused videos. More results are provided in the supplementary material.

Intra-Feature Pearson Constraint	Features			t2v Retrieval		
	v_{L_1}	v_{L_2}	v_{L_3}	R@1	R@5	Mean
No	✓			42.8	71.6	65.2
Yes	✓			43.6	72.0	65.7
No	✓	✓		47.2	73.2	67.5
Yes	✓	✓		47.3	73.2	67.8
No	✓	✓	✓	47.8	73.0	67.8
Yes	✓	✓	✓	**47.8**	**74.1**	**68.7**

Table 4: Ablation study of Pearson Constraint in EERCF on MSRVTT-1K-Test dataset.

Reranking Method		Performance			Efficiency
		R@1	R@5	Mean	(FLOPs)
Text Supervised	X-CLIP	46.8	72.8	66.9	11.5k
	DRL	45.2	73.1	67.4	11.6k
	EERCF(v_{L_2})	**47.3**	**73.2**	**67.8**	**0.8k**
Unsupervised	Video Sim	42.5	68.7	63.2	1.8k

Table 5: Ablation study of different re-ranking methods on MSRVTT-1K-Test dataset.

diverse query texts. It also indicates the indispensability of the TIB, responsible for extracting fine-grained adaptive features in the second re-ranking stage.

Top-k re-ranking hyper-parameter. As shown in Fig.3, EERCF obtains stable retrieval performance improvement when top-k ranges from 10 to 50. We also observe that there is a slight decrease in retrieval performance when the top-k increases to 100. We believe this is due to the re-ranking stage excessively focusing on visual local details and the noise introduced by increasing the top-k can cause more disturbances to the re-ranking stage. And there is some preliminary evidence that X-CLIP also benefits from two-stage retrieval, as the original X-CLIP essentially re-ranking all videos, whereas our experiments only re-ranking the top-50 videos as shown in Tab.2 and Tab.5. Also with Pearson Constraint, we have observed consistent performance improvement across different top-k values in retrieval.

Intra-Feature Pearson Constraint. We also demonstrate the effectiveness of the Intra-Feature Pearson Constrain through experiments. As shown in Tab.4, the retrieval performance has shown consistent improvement through the use of the Intra-Feature Pearson Constrain from coarse to fine video features. The above-mentioned results have proven the effectiveness of using the Pearson Constrain to enhance the uniqueness of features and reduce redundant information.

Different Re-ranking Methods. Our efficiency improvement is primarily due to the second stage of re-ranking. So we evaluate our method against previously computationally expensive methods in the second stage. We also compare with a commonly used unsupervised re-ranking method named Video Sim(Wei et al. 2020). Each video in the candidate set undergoes a voting process based on inter-video similarity, with the resulting vote count determining the final re-ranking order. To clarify, we perform re-ranking on the same candidate set. As shown in Tab. 5, our method outperforms the alternatives in both performance and efficiency.

Visualization of Coarse-to-Fine Retrieval. We illustrate the coarse-to-fine reranking in Fig.4, which shows text-agnostic video features often find confused results, and TIB can gradually dig relevant frame-level or patch-level cues for accurate results. We observe that fine-grained information can lead to more precise matching with query texts, but inevitably introduces some noise. This also highlights the necessity of two-stage retrieval, which strikes a good balance between overly abstract and detailed video representations.

Conclusion

This paper develops a novel EERCF framework for efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. To this end, a parameter-free text-gated interaction block is exploited to fine-grained video representations. At the same time, we use a Pearson coefficient trick to optimize representation learning. Finally, using a coarse-to-fine retrieval strategy, our approach achieves the best trade-off between performance and cost on all popular datasets.

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 1728–1738.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 190–200.

Conde, M. V.; and Turgutlu, K. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *CVPR*, 3956–3960.

Croitoru, I.; Bogolin, S.-V.; Leordeanu, M.; Jin, H.; Zisserman, A.; Albanie, S.; and Liu, Y. 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, 11583–11593.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *Arxiv*.

Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qie, X.; and Luo, P. 2022. Bridging video-text retrieval with multiple choice questions. In *CVPR*, 16167–16176.

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 *CVPR*, 5006–5015.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970. IEEE.

Hu, F.; Chen, A.; Wang, Z.; Zhou, F.; Dong, J.; and Li, X. 2022a. Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, 444–461.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022b. Scaling up vision-language pre-training for image captioning. In *CVPR*, 17980–17989.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *arXiv*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900.

Li, X.; Xu, C.; Yang, G.; Chen, Z.; and Dong, J. 2019. W2vv++ fully deep learning for ad-hoc video search. In *ACMMM*, 1786–1794.

Li, X.; Zhou, F.; Xu, C.; Ji, J.; and Yang, G. 2020. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23: 4351–4362.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, 11915–11925.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 319–335.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *ACMMM*, 638–647.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv*.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022. Disentangled representation learning for text-video retrieval. *arXiv*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 4581–4591.

Wei, W.; Jiang, M.; Zhang, X.; Liu, H.; and Tian, C. 2020. Boosting cross-modal retrieval With MVSE++ and reciprocal neighbors. *IEEE Access*, 8: 84642–84651.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 5036–5045.

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2023. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment. *ICML*.

Zhao, S.; Zhu, L.; Wang, X.; and Yang, Y. 2022. Center-CLIP: Token Clustering for Efficient Text-Video Retrieval. In *SIGIR*.