This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EagleNet: Energy-Aware Fine-Grained Relationship Learning Network for Text-Video Retrieval

Yuhan Chen[1] Pengwen Dai[2,3*] Chuan Wang[4] Dayan Wu[5] Xiaochun Cao[2,3*]
[1]Sun Yat-sen University [2]Shenzhen Campus of Sun Yat-sen University
[3]Shenzhen Key Laboratory of Adversarial Artificial Intelligence
[4]Beijing Jiaotong University [5]Institute of Information Engineering, CAS
Guangzhou, China[1] Shenzhen, China[2,3] Beijing, China[4,5]
draym28@gmail.com daipw@mail.sysu.edu.cn wangchuan@bjtu.edu.cn
wudayan@iie.ac.cn caoxiaochun@mail.sysu.edu.cn

Abstract

Text-video retrieval tasks have seen significant improvements due to the recent development of large-scale vision-language pre-trained models. Traditional methods primarily focus on video representations or cross-modal alignment, while recent works shift toward enriching text expressiveness to better match the rich semantics in videos. However, these methods use only interactions between text and frames/video, and ignore rich interactions among the internal frames within a video, so the final expanded text cannot capture frame contextual information, leading to disparities between text and video. In response, we introduce Energy-Aware Fine-Grained Relationship Learning Network (EagleNet) to generate accurate and context-aware enriched text embeddings. Specifically, the proposed Fine-Grained Relationship Learning mechanism (FRL) first constructs a text-frame graph by the generated text candidates and frames, then learns relationships among texts and frames, which are finally used to aggregate text candidates into an enriched text embedding that incorporates frame contextual information. To further improve fine-grained relationship learning in FRL, we design Energy-Aware Matching (EAM) to model the energy of text-frame interactions and thus accurately capture the distribution of real text-video pairs. Moreover, for more effective cross-modal alignment and stable training, we replace the conventional softmax-based contrastive loss with the sigmoid loss. Extensive experiments have demonstrated the superiority of EagleNet across MSRVTT, DiDeMo, MSVD, and VATEX.

1. Introduction

As the volume of video data being generated and stored continues to grow, the need for effective content retrieval models has become more urgent than ever. Text-Video Retrieval (TVR) has been extensively studied and shows great progress in recent years. In Retrieval-Augmented Generation (RAG) [15], TVR improves the accuracy of generative models by integrating video content into knowledge-based systems, enhancing user interactions in applications like dialogue systems.

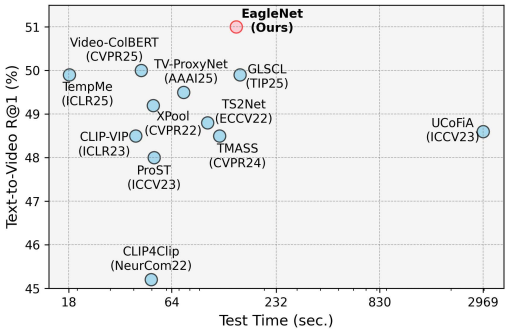


Figure 1. The performance-efficiency comparison among SOTA models (ViT-B/16) on MSRVTT datasets.

Early works [5, 9, 10, 13, 14, 16, 26, 29, 31, 46, 49, 55, 58] mainly rely on pre-extracted features, which are learned by independent text/image encoders. Additionally, due to the huge development in Graph Neural Network (GNN), many researchers utilize graph learning techniques in text-video retrieval tasks. As the groundbreaker, HGR [5] constructs a semantic graph for each sentence, and utilizes GNN to learn text embeddings. Inspired by HGR, many follow-up works [13, 19, 23, 44] employ GNN for text-video structural relationship learning. For example, ACP [13] introduces knowledge graph learning model to help complete semantic information for retrieval. However, all of these works are based on data pre-processing approaches with heavy external knowledge or outdated encoders, limiting the flexibility and increasing the complexity of the model.

*Corresponding authors: Pengwen Dai (daipw@mail.sysu.edu.cn) and Xiaochun Cao (caoxiaochun@mail.sysu.edu.cn).

Instead, recent studies incorporate the emergent large-scale image-text pre-trained models into text-video retrieval tasks [3, 7, 12, 17, 27, 32–34, 38, 41, 43, 48, 51, 53, 54, 57]. They utilize CLIP image and text encoder [37] as multi-modal encoders to achieve strong performance on many text-video retrieval benchmarks. CLIP4Clip [33], the first work that applies CLIP in this task, proposes using a temporal aggregator to fuse frames and then align text and video features. Later, many works [3, 7, 12, 17, 32, 41, 57] focus on obtaining high-quality video features to enhance cross-modal matching. Besides, some other works [27, 38, 48, 54, 57, 59] propose to improve the text-video alignment, such as conducting matching at different granularities. However, most models overlook the insufficient expressiveness of text embeddings. The brevity of video descriptions often prevents text embeddings from fully reflecting the underlying video semantics. To fill this gap, TMASS [43] stochastically samples enriched text within a flexible range, the radius of which is determined by the similarity of text and frame embeddings. Similarly, TV-ProxyNet [51] refines the text query into a specific text proxy for each candidate video proxy through video-aware directors and text-video similarities. Nevertheless, these methods focus solely on interactions between text and frames/video when expanding text semantics, and ignore relationships among video frames, so the final enriched text cannot capture frame contextual information, leading to disparities between text and video.

We argue that the internal frame-to-frame relations within the video should also be incorporated when expanding text semantics. This enables the model to capture both global and temporal semantics of videos, allowing the generated text embeddings to reflect not only the semantics of individual frames but also their contextual dependencies. Moreover, by establishing relationships among video semantic units (frames), the model can effectively identify and suppress redundant information and noise, leading to more accurate, contextually enriched text embeddings.

To this end, we introduce the **E**nergy-**A**ware Fine-**G**rained Relationship **Le**arning **Net**work (**EagleNet**), a framework designed to generate accurate and context-aware enriched text embeddings by exploring fine-grained relationships among text and video frames (see Fig. 2). The model consists of two key components: 1) Fine-Grained Relationship Learning (FRL) and 2) Energy-Aware Matching (EAM). Specifically, FRL first generates multiple text candidates using a stochastic text modeling strategy, and constructs a text-frame relational graph. We then design an effective relational graph attention network to learn the inter-connections among texts and frames. These connections incorporate both text-frame interactions and frame contextual information, which are subsequently used to aggregate the text candidates into the final enriched text embedding. To further enhance fine-grained text-frame relationship learning in FRL, EAM employs energy-based model to accurately model the energy of detailed text-frame interactions and thus capture the true distribution of text-video pairs, improving the overall matching performance. Moreover, we proposed replacing the conventional softmax-based contrastive loss [35] with sigmoid loss [56] for more effective cross-modal alignment and stable training. Softmax loss normalizes the logits across both dimensions of the batch similarity matrix, making it sensitive to the negative samples and batch size [8, 24, 56]. Instead, sigmoid loss treats each pair independently, so it can mitigate these drawbacks and is naturally suitable for the multi-matched scenario of TVR, where one text may be semantically similar to multiple videos or vice versa. Extensive experiments have been conducted on various benchmarks, demonstrating our effectiveness and superiority.

Our main contributions can be summarized as follows:

- We design a fine-grained relationship learning network that considers not only text-video interaction but also internal frame-to-frame relations within videos, generating accurate enriched text embedding that incorporates frames' contextual information;
- To the best of our knowledge, we are the first to integrate EBM into TVR, which effectively models fine-grained text-frame interactions, enhancing the relational learning and cross-modal alignment.
- Extensive experiments illustrate the superiority of EagleNet and the effectiveness of each component, showing improvement over state-of-the-art models across four benchmark datasets.

2. Related Work

Text-Video Retrieval. Text-Video Retrieval (TVR) needs to align text embeddings with video embeddings [18, 21, 47, 50]. Early works are mostly offline models that focus on designing a matching algorithm for pre-extracted features [5, 9, 10, 13, 14, 16, 26, 29, 31, 46, 49, 55, 58]. JPoSE [49] and DualEncoding [10] extract video-text features offline, then perform embedding alignment in a common space. To enhance representation quality, CE [31], MMT [14], and HiT [29] employ multiple unimodal experts (OCR, speech, etc.) to enrich video features. TeachText [9] adopts multiple text encoders to gather text information and obtain more credible alignments. Besides, HGR [5] and T2vlad [46] build semantic graphs or shared centers to explore global-local relations. However, these models employ additional analysis tools or experts to pre-process data, which increases the complexity and limits flexibility.

Recently, more and more researchers aim to seek end-to-end solutions [7, 12, 16, 17, 26, 27, 32–34, 43, 48, 53, 58]. ClipBERT [26] and Frozen [2] first propose efficient end-to-end pre-trained schemes. MCQ [16] builds cross-modal associations by predicting verb or noun features. Later, benefiting from the large-scale vision-language pre-trained mod-

els [37], CLIP4clip [33] uses CLIP as the backbone network, significantly improving TVR performance, and inspiring a series of works [3, 12, 17, 22, 32, 40, 58]. Centerclip [58] conducts segment-level clustering, which reduces token redundancy and computational overhead. TS2Net [32] designs the token shift module to realize the perception of local movement between frames. CLIP-VIP [53] learns video proxies to incorporate frame context. XPool [17] uses text as a condition to guide the aggregation of video tokens. UCoFiA [48] conducts hierarchical alignment to achieve coarse- and fine-grained matching. Similarly, GLSCL [57] proposes a global and local interaction module for coarse- and fine-grained semantic consistent learning. ProST [27] extracts higher-level prototypes for texts and videos and conducts spatio-temporal matching. TempMe [41] introduces a multi-granularity framework to progressively combine neighboring clips, reducing video redundancy and enhancing temporal modeling across frames. Video-ColBERT [38] utilizes spatiotemporal token interactions, learnable query and visual expansion, and sigmoid loss training.

However, most existing works focus on learning high-quality video embeddings or improving multi-modal alignment, while overlooking the limited expressiveness of textual representations. TMASS [43] extends text embeddings effectively by sampling stochastic text embeddings within a flexible range, whose radius is determined by the similarity between texts and videos. Similarly, TV-ProxyNet [51] converts the text query to a specific text proxy for each candidate video proxy through video-aware directors and text-video similarities. Nevertheless, these methods only consider text-video interactions while expanding text semantics, ignoring the internal frame-to-frame relations within a video, so the final enriched text cannot capture frame context, leading to disparities between text and video.

Graph Neural Networks. Graph Neural Network (GNN) has received significant interest in recent years due to its powerful ability in real-world applications based on graph-structured data [25, 39, 42]. The Graph Attention Network (GAT) [42] is the first to introduce the attention mechanism into GNN, and it shows great potential in modeling relationships. Furthermore, to tackle complex real-life problems related to heterogeneous graphs, such as processing knowledge graphs, Relational GNN (RGNN) [39] and many successors have been proposed. HGR [5] is the first work to employ GNN in TVR; it constructs semantic graphs for texts by grammar and extracts representations using GNN. Later, many following works are proposed to employ GNN in TVR [19, 23, 44]. ACP [13] uses knowledge graphs as prior knowledge to complete the semantic information. SHE-Net [54] introduces CLIP as the backbone and uses hierarchical text graphs to guide the aggregation of temporal frames and spatial regions. But these works are based on pre-extracted features and heavily rely on external knowledge, limiting the flexibility and increasing model complexity.

Unlike previous methods, our approach captures both the text-video interactions and rich relationships among video frames, which are essential for comprehending the video context. By leveraging these fine-grained interactions, the model is able to generate high-quality, context-aware text embeddings and achieve more accurate text-video alignment.

3. Method

3.1. Preliminary

We first introduce the standard pipeline of the text-video retrieval task, including: 1) obtaining representations of text and video; 2) training and inference.

3.1.1. Text and Video Representations

Text-video retrieval task aims to align the text and video in a joint space. The first step is to learn the text and video embeddings using the vision-language pre-trained model CLIP [37]. Specifically, given a text-video pair $\{t, v\}$, TVR first samples M frames from v uniformly, $[f_1, f_2, \cdots, f_M]$, and then feeds them into CLIP text encoder $\phi_{\text{t}}(\cdot)$ and image encoder $\phi_{\text{v}}(\cdot)$, obtaining text embedding $\mathbf{t}$ and frame embeddings $\mathbf{F}$ as:

$$\begin{aligned} \mathbf{t} &= \phi_{\text{t}}(t) \in \mathbb{R}^d, \\ \mathbf{F} &= \phi_{\text{v}}([f_1, f_2, \cdots, f_M]) = [\mathbf{f}_1 \| \mathbf{f}_2 \| \cdots \| \mathbf{f}_M]^\top \in \mathbb{R}^{M \times d}, \end{aligned} \tag{1}$$

where "$\|$" is the column concatenation operation, and d is the hidden dimension of embeddings.

Then a frame fusion module is employed to obtain the global video feature $\mathbf{v} = \pi(\mathbf{F}) \in \mathbb{R}^d$, to reduce the temporal redundancy in video and align $\mathbf{t}$ with $\mathbf{v}$. A simple solution is average pooling. But it treats all the frames equally, overlooking the important frames and retaining noise information from irrelevant frames. We follow XPool [17] (Appendix A) to use $\mathbf{t}$ as condition, such that the fusion module can attach more attention to the text-related frames:

$$\mathbf{v} = \pi(\mathbf{F}|\mathbf{t}) \in \mathbb{R}^d, \tag{2}$$

which uses cross-attention to fuse frames guided by text.

3.1.2. Training and Inference

In training stage, for each training batch $\{(t_i, v_i)\}_{i=1}^B$ with batch size B, a symmetric cross-entropy loss $\mathcal{L}_{\text{ce}}$ is applied to minimize the distance of relevant text-video pairs while maximizing that between irrelevant pairs, similar to contrastive learning [35]. We consider both text-to-video retrieval ($\mathcal{L}_{\text{ce}}^{t2v}$) and video-to-text retrieval ($\mathcal{L}_{\text{ce}}^{v2t}$):

$$\begin{aligned} \mathcal{L}_{\text{ce}}(\mathbf{t}, \mathbf{v}) &= \frac{1}{2}\Big(\mathcal{L}_{\text{ce}}^{t2v}(\mathbf{t}, \mathbf{v}) + \mathcal{L}_{\text{ce}}^{v2t}(\mathbf{t}, \mathbf{v})\Big), \\ \mathcal{L}_{\text{ce}}^{t2v}(\mathbf{t}, \mathbf{v}) &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{\text{s}(\mathbf{t}_i, \mathbf{v}_i)\cdot\lambda}}{\sum_{j=1}^{B} e^{\text{s}(\mathbf{t}_i, \mathbf{v}_j)\cdot\lambda}}, \\ \mathcal{L}_{\text{ce}}^{v2t}(\mathbf{t}, \mathbf{v}) &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{\text{s}(\mathbf{t}_i, \mathbf{v}_i)\cdot\lambda}}{\sum_{j=1}^{B} e^{\text{s}(\mathbf{t}_j, \mathbf{v}_i)\cdot\lambda}}, \end{aligned} \tag{3}$$

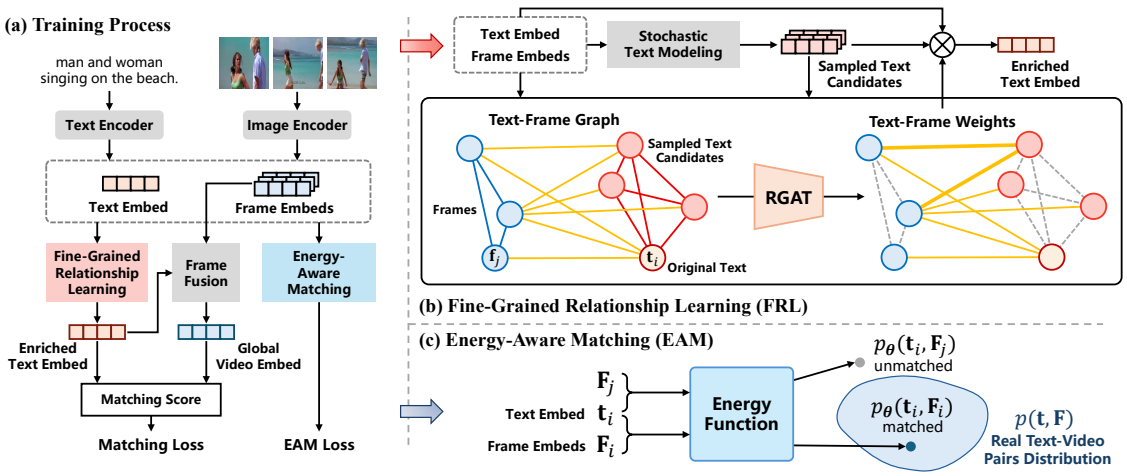


Figure 2. Diagram of EagleNet. (a) Overview of the training process. (b) Fine-Grained Relationship Learning (FRL) first samples multiple text candidates, then constructs a text-frame graph to learn both text-frame and frame-frame relationships, which are then used to aggregate text candidates into a final enriched text embedding aware of video context information. (c) Energy-Aware Matching (EAM) improves the relationships learning in FRL by capturing detailed text-frame interactions and accurately models true text-video pairs distribution, thereby also enhancing the final matching performance from a fine-grained perspective.

where $\mathrm{s}(\cdot,\cdot)$ is similarity function (*e.g.*, cosine similarity) and λ is a learnable factor.

In inference, given n_t texts $\{t_i\}_{i=1}^{n_t}$ and n_v video clips $\{v_j\}_{j=1}^{n_v}$, for a query text t_{query}, the matched video clip v^* is the one that has the highest matching score and vice versa:

$$v^* = \arg\max_j \mathrm{s}(\mathbf{t}_{\text{query}}, \mathbf{v}_j), \quad t^* = \arg\max_i \mathrm{s}(\mathbf{t}_i, \mathbf{v}_{\text{query}}).$$

3.2. Our Proposed Model: EagleNet

In this section we introduce **EagleNet** (Fig. 2), which comprises two key components: 1) Fine-Grained Relationship Learning (**FRL**) and 2) Energy-Aware Matching (**EAM**). First, FRL generates context-aware enriched text by sampling text candidates and extracting fine-grained relationships among texts and frames (Sec. 3.2.1). Then, EAM enhances FRL in learning text-frame interactions and improves the final matching from a fine-grained view (Sec. 3.2.2). Finally, the sigmoid function is employed in the final loss for more effective alignment and stable training (Sec. 3.2.3).

3.2.1. Fine-Grained Relationship Learning

Existing models only consider text-frame interactions when expanding text semantics. However, the internal frame-frame relations are also important to capture both global and temporal semantics. Specifically, we first sample multiple text candidates, and the texts and frames are treated as nodes to form a text-frame graph, which includes both text-frame and frame-frame relations. Then the relations are extracted by relational graph learning and used to fuse text candidates, generating text embedding that reflects not only the semantics of individual frames but also their contextual dependencies.

Text-Frame Graph Construction. Given text embedding $\mathbf{t}$ and frame embeddings $\mathbf{F} = [\mathbf{f}_1 \| \mathbf{f}_2 \| \cdots \| \mathbf{f}_M]^\top$, we first sample S stochastic text candidates $\{\mathbf{t}_i^{\text{sto}}\}_{i=1}^{S}$ using stochastic text modeling [43] (Appendix B), then concatenate texts and frames to get the node matrix as:

$$\begin{aligned} \mathbf{X} &= \left[\mathbf{t} \| \mathbf{t}_1^{\text{sto}} \| \mathbf{t}_2^{\text{sto}} \| \cdots \| \mathbf{t}_S^{\text{sto}} \| \mathbf{f}_1' \| \mathbf{f}_2' \| \cdots \| \mathbf{f}_M' \right]^\top \in \mathbb{R}^{n \times d}, \\ \mathbf{f}_j' &= \mathbf{f}_j + \mathrm{PE}(j), \end{aligned} \tag{4}$$

where $n = 1 + S + M$ is the number of nodes, and $\mathrm{PE}(\cdot)$ is learnable temporal positional embedding on frames. Thus, there are three types of edges (relations): text-text, frame-frame, and text-frame, denoted by $r \in \{tt, ff, tf\}$.

Relationship Learning. To extract useful information from the complex relations among texts and frames, we propose a simple yet effective version of the Relational Graph Attention Network (**RGAT**) with L layers. The layer index is omitted here for brevity. In each layer, for any node pair (i, j) linked by relation r, the edge weight is given by

$$e_{ij}^{r,h} = \psi^r\left(\left[\mathbf{W}^{r,h}\mathbf{h}_i \| \mathbf{W}^{r,h}\mathbf{h}_j\right]\right) \in \mathbb{R}, \tag{5}$$

where $h = 1, 2, \cdots, H$ is the indices of attention head, $\mathbf{h}_i$ is the feature of node i (in the first layer $\mathbf{h}_i$ is the i-th row of $\mathbf{X}$), $\mathbf{W}^{r,h}$ is learnable matrix, and $\psi^r(\cdot) : \mathbb{R}^{2d} \mapsto \mathbb{R}$ is projector to obtain edge weights. Then, the attention score is obtained as,

$$\alpha_{ij}^{r,h} = \frac{\exp\left(\mathrm{LeakyReLU}(e_{ij}^{r,h})\right)}{\sum_{k \in \mathcal{N}_i^r} \exp\left(\mathrm{LeakyReLU}(e_{ik}^{r,h})\right)}, \tag{6}$$

where $\mathcal{N}_i^r$ is the neighbor set of node i linked by relation r. Finally, the output node representation of this layer $\mathbf{h}_i'$ is

$$\mathbf{h}_i' = \sigma\Big(\mathbf{W}_{\text{out}}\mathbf{h}_i + \Big\|_{h=1}^{H} \sum_r \sum_{j \in \mathcal{N}_i^r} \alpha_{ij}^{r,h} \mathbf{W}^{r,h}\mathbf{h}_j\Big), \tag{7}$$

where $\sigma(\cdot)$ is ReLU, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{Hd\times d}$ is learnable weights for residual connection. Then $\mathbf{h}'_i$ is input of the next layer. Matrix form and more details can be found in Appendix C.

Note that here $\mathbf{h}'_i$ only serves as an intermediate. The target of FRL is text-frame relations e^{tf}_{ij}, which are regarded as the importance of text embeddings for the frames. These weights are then used to aggregate text embeddings into a final enriched text. Specifically, the text-frame edge weights from **the final layer** are averaged across attention heads to get the final text-frame weights, *i.e.*,

$$e^{tf}_{ij} = \frac{1}{H}\sum_{h=1}^{H} e^{tf,h}_{ij}, \quad \begin{array}{l} i \in [1, 1+S], \\ j \in [1+S+1, 1+S+M], \end{array} \tag{8}$$

then we average and normalize the text-frame weights to obtain the attention score of each text node:

$$w_i = \frac{\exp(e^{tf}_i)}{\sum_{k=1}^{1+S}\exp(e^{tf}_k)}, \quad e^{tf}_i = \frac{1}{M}\sum_{j=1+S+1}^{1+S+M} e^{tf}_{ij}, \tag{9}$$

so the final enriched text embedding is

$$\mathbf{t}^{\text{gen}} = \sum_{i=1}^{1+S} w_i \mathbf{X}_i = w_1\mathbf{t} + \sum_{i=1}^{S} w_{i+1}\mathbf{t}^{\text{sto}}_i \in \mathbb{R}^d. \tag{10}$$

$\mathbf{t}^{\text{gen}}$ is used as the condition for frame fusion (Eq. (2)) to obtain video embedding $\mathbf{v}$, then $\mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{gen}}, \mathbf{v})$ is calculated (Eq. (3)). Following [43] (Appendix B), we further obtain a support text $\mathbf{t}^{\text{sup}}$, and calculate the support loss $\mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{sup}}, \mathbf{v})$. $\mathbf{t}^{\text{sup}}$ serves as the boundary of the stochastic modeling range, within which $\mathbf{t}^{\text{sto}}_i$ is sampled. So the matching loss is

$$\mathcal{L}_{\text{match}} = \mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{gen}}, \mathbf{v}) + \lambda_{\text{sup}}\mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{sup}}, \mathbf{v}), \tag{11}$$

where λ_{sup} is a coefficient.

3.2.2. Energy-Aware Matching

Since $\mathcal{L}_{\text{match}}$ only aligns text and video at global level, we propose EAM to assist FRL in learning fine-grained text-frame relations. EAM employs energy-based model to effectively capture data distributions [6, 11, 30], thereby enhancing the matching from a fine-grained perspective.

To be specific, given text embedding $\mathbf{t}$ and frame embeddings $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_M]$, EAM models the real text-video pair distribution $p(\mathbf{t}, \mathbf{F})$ using Boltzmann distribution $p_\theta(\mathbf{t}, \mathbf{F}) = \frac{\exp(-E_\theta(\mathbf{t},\mathbf{F}))}{Z_\theta}$, where Z_θ is the normalizing constant, θ is learnable parameter, and $E_\theta(\mathbf{t}, \mathbf{F}) \in \mathbb{R}$ denotes the text-video energy. We first calculate the energy for each text-frame pair and then average them to obtain the text-video energy $E_\theta(\mathbf{t}, \mathbf{F}) = \frac{1}{M}\sum_i^M E_\theta(\mathbf{t}, \mathbf{f}_i)$. Such that EAM can fully utilize the information in fine-grained text-frame interactions. Simultaneously, E_θ assigns low energy scores to the true text-video pairs and high energy to mismatched ones, further assisting the model in achieving more precise text-video alignment.

Additionally, besides averaging text-frame energies (`Avgpool`), other pooling approaches, *e.g.*, `Maxpool`, are also options. This will be discussed in Sec. 4.3.3.

Energy Functions. Since the energy score indicates the relevance of a text-video pair, we can adopt a similarity function as E_θ, such as negative cosine similarity (`CosSim`):

$$E(\mathbf{t}, \mathbf{F}) = \frac{1}{M}\sum_{i=1}^{M} E(\mathbf{t}, \mathbf{f}_i) = -\frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{t}^\top \mathbf{f}_i}{\|\mathbf{t}\|\|\mathbf{f}_i\|}. \tag{12}$$

Or we can use the bilinear scoring function with learnable parameters $\mathbf{W}$ (`Bilinear`):

$$E_\theta(\mathbf{t}, \mathbf{F}) = -\frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{t}^\top \mathbf{W}\mathbf{f}_i}{\|\mathbf{t}\|\|\mathbf{f}_i\|}. \tag{13}$$

For greater flexibility, MLP is also a strong choice (`MLP`):

$$E_\theta(\mathbf{t}, \mathbf{F}) = \frac{1}{M}\sum_{i=1}^{M}\text{MLP}([\mathbf{t}\|\mathbf{f}_i]). \tag{14}$$

Training EAM. In practice, EAM is trained by $\mathcal{L}_{\text{eam}} = -\mathbb{E}_{\{\mathbf{t},\mathbf{F}\}\sim p(\mathbf{t},\mathbf{F})}[\log p_\theta(\mathbf{t}, \mathbf{F})]$, this negative log-likelihood loss can be reformulated as:

$$\mathcal{L}_{\text{eam}} \approx \frac{1}{B}\sum_{i=1}^{B} E_\theta(\mathbf{t}_i, \mathbf{F}_i) - \frac{1}{B}\sum_{j=1}^{B} E_\theta(\tilde{\mathbf{t}}_j, \tilde{\mathbf{F}}_j), \tag{15}$$

where $\{(\tilde{\mathbf{t}}_j, \tilde{\mathbf{F}}_j)\}_{j=1}^B$ are fake text-video pairs sampled from $p_\theta(\tilde{\mathbf{t}}, \tilde{\mathbf{F}})$ using a K-step Markov chain Monte Carlo (MCMC) sampling:

$$\begin{cases} \tilde{\mathbf{t}}_j^{(k+1)} = \tilde{\mathbf{t}}_j^{(k)} - \eta\nabla_{\tilde{\mathbf{t}}_j^{(k)}} E_\theta(\tilde{\mathbf{t}}_j^{(k)}, \tilde{\mathbf{F}}_j^{(k)}) + \boldsymbol{\epsilon}_{t_j}^{(k+1)}, \\ \tilde{\mathbf{F}}_j^{(k+1)} = \tilde{\mathbf{F}}_j^{(k)} - \eta\nabla_{\tilde{\mathbf{F}}_j^{(k)}} E_\theta(\tilde{\mathbf{t}}_j^{(k)}, \tilde{\mathbf{F}}_j^{(k)}) + \boldsymbol{\epsilon}_{v_j}^{(k+1)}, \end{cases} \tag{16}$$

where $\tilde{\mathbf{t}}_j^{(0)} \in \mathbb{R}^d$ and $\tilde{\mathbf{F}}_j^{(0)} \in \mathbb{R}^{M\times d}$ are given initial samples, η is the step size, and $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(0, \sigma^2)$ with a given noise σ^2. We stop the gradients of $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{F}}$ after sampling. More details of the derivation and sampling algorithm can be found in Appendix D. EAM is only involved in the training process, and thus will not increase inference cost.

We incorporate $\mathcal{L}_{\text{eam}}$ to define the total objective:

$$\begin{aligned} \mathcal{L}_{\text{total}} =& \mathcal{L}_{\text{match}} + \lambda_{\text{eam}}\mathcal{L}_{\text{eam}} \\ =& \mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{gen}}, \mathbf{v}) + \lambda_{\text{sup}}\mathcal{L}_{\text{ce}}(\mathbf{t}^{\text{sup}}, \mathbf{v}) + \lambda_{\text{eam}}\mathcal{L}_{\text{eam}}, \end{aligned} \tag{17}$$

where λ_{eam} is a coefficient.

3.2.3. Sigmoid Loss

The softmax-based contrastive loss ($\mathcal{L}_{\text{ce}}$) is sensitive to negative samples and batch size due to the batch normalization [8, 24, 56]. Instead, the sigmoid loss [56] models each text-video pair independently, enabling more effective alignment and stable optimization, so is naturally suitable for the multi-matched scenario of TVR, where one text may be semantically similar to multiple videos or vice versa.

$$\mathcal{L}_{\text{sig}}(\mathbf{t}, \mathbf{v}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\log\frac{1}{1+e^{\mathbb{I}_{ij}(\tau\cdot\text{s}(\mathbf{t}_i,\mathbf{v}_j)+b)}} \tag{18}$$

	Method	Text-to-Video R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Video-to-Text R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Rsum↑
	Not End-to-End Method											
	CE [31] BMVC19	20.6	50.3	64.0	5.3	-	20.9	48.8	62.4	6.0	-	267.0
	MMT [14] ECCV20	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3	307.5
	T2vlad [46] CVPR21	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	-	321.5
	HiT [29] ICCV21	27.7	59.2	72.0	2.9	-	28.8	60.3	72.3	3.0	-	320.3
	TeachText [9] ICCV21	29.6	61.6	74.2	3.0	-	32.1	62.7	75.0	3.0	-	335.2
	End-to-End Method											
non-CLIP	ClipBERT [26] CVPR21	22.0	46.8	59.9	6.0	-	-	-	-	-	-	-
non-CLIP	SupportSet [36] ICLR21	30.1	58.5	69.3	3.0	-	28.5	58.6	71.6	3.0	-	316.6
non-CLIP	Frozen [2] ICCV21	31.0	59.5	70.5	3.0	-	-	-	-	-	-	-
non-CLIP	BridgeFormer [16] CVPR22	37.6	64.8	75.1	3.0	-	-	-	-	-	-	-
non-CLIP	TVMM [28] NeurIPS22	36.2	64.2	75.7	3.0	-	34.8	63.8	73.7	3.0	-	348.4
ViT-B/32	CLIP4Clip* [33] NeurCom22	43.2	71.4	80.6	**2.0**	16.5	42.5	70.9	81.0	**2.0**	12.2	389.6
ViT-B/32	XPool†* [17] CVPR22	45.8	73.8	82.0	**2.0**	14.0	**45.7**	73.1	82.3	**2.0**	9.4	402.7
ViT-B/32	TS2Net* [32] ECCV22	45.4	72.9	82.9	**2.0**	14.2	44.8	72.6	81.1	**2.0**	10.3	399.7
ViT-B/32	CLIP-VIP* [53] ICLR23	46.0	72.6	82.2	**2.0**	14.4	43.6	71.5	82.3	**2.0**	10.8	398.2
ViT-B/32	ProST* [27] ICCV23	46.3	73.6	82.4	**2.0**	12.9	45.0	72.9	82.2	**2.0**	**8.9**	402.4
ViT-B/32	TMASS†* [43] CVPR24	43.4	73.2	82.1	**2.0**	14.3	45.0	73.2	82.7	**2.0**	9.5	399.6
ViT-B/32	RAP [3] ACL24	44.8	71.4	81.5	-	14.4	44.0	71.9	82.4	-	10.1	396.0
ViT-B/32	TFVL [40] TNNLS25	45.2	70.4	80.1	-	15.2	45.5	71.5	81.0	-	11.0	393.7
ViT-B/32	SHE-Net [54] TCSVT25	45.3	**74.9**	**84.2**	**2.0**	**12.5**	44.8	**74.8**	**83.7**	**2.0**	9.1	407.7
ViT-B/32	GLSCL* [57] TIP25	**47.2**	73.5	82.7	**2.0**	13.1	45.6	73.5	82.7	**2.0**	9.8	405.2
ViT-B/32	TempMe* [41] ICLR25	43.3	71.2	81.2	**2.0**	15.3	45.0	71.5	80.8	**2.0**	11.0	393.0
ViT-B/32	TV-ProxyNet†* [51] AAAI25	46.7	73.0	81.8	**2.0**	14.5	45.4	71.1	82.4	**2.0**	11.0	400.4
ViT-B/32	Video-ColBERT†* [38] CVPR25	45.9	71.6	82.7	**2.0**	13.7	42.1	70.4	81.1	**2.0**	10.8	393.8
ViT-B/32	**EagleNet (Ours)**	**47.2**	74.8	84.0	**2.0**	14.4	**45.7**	73.7	83.4	**2.0**	9.4	**408.8**
ViT-B/16	CLIP4Clip* [33] NeurCom22	45.2	72.2	81.4	2.0	15.0	42.9	70.9	80.6	**2.0**	11.6	393.2
ViT-B/16	XPool†* [17] CVPR22	49.2	73.9	82.6	2.0	13.6	48.0	74.9	82.9	**2.0**	9.7	411.5
ViT-B/16	TS2Net* [32] ECCV22	48.8	75.0	85.0	2.0	13.0	46.4	75.4	85.3	**2.0**	8.6	415.9
ViT-B/16	CLIP-VIP* [53] ICLR23	48.5	75.5	83.9	2.0	12.5	47.6	74.8	85.0	**2.0**	9.6	415.3
ViT-B/16	UCoFiA* [48] ICCV23	48.6	75.5	83.8	2.0	**11.7**	49.0	75.7	84.2	**2.0**	10.0	416.8
ViT-B/16	ProST* [27] ICCV23	48.0	74.7	84.0	2.0	12.7	48.3	75.2	84.5	**2.0**	8.7	414.7
ViT-B/16	TMASS†* [43] CVPR24	48.5	74.1	83.4	2.0	14.4	46.7	75.9	85.6	**2.0**	9.1	414.2
ViT-B/16	RAP [3] ACL24	46.5	73.9	82.0	-	12.1	45.3	76.4	84.8	-	9.1	408.9
ViT-B/16	TFVL [40] TNNLS25	48.7	73.5	83.0	-	13.4	47.3	75.3	83.7	-	9.5	411.5
ViT-B/16	GLSCL* [57] TIP25	49.9	**76.3**	84.1	2.0	11.9	48.3	75.8	84.6	**2.0**	**8.1**	419.0
ViT-B/16	TempMe* [41] ICLR25	49.9	75.3	84.8	2.0	12.1	47.5	75.2	85.1	**2.0**	8.8	417.8
ViT-B/16	TV-ProxyNet†* [51] AAAI25	49.5	74.9	83.4	2.0	12.6	47.9	74.8	84.1	**2.0**	9.4	414.6
ViT-B/16	Video-ColBERT†* [38] CVPR25	50.0	**76.3**	84.3	1.5	13.1	47.9	74.4	84.9	**2.0**	9.0	417.8
ViT-B/16	**EagleNet (Ours)**	**51.0**	76.2	**85.6**	**1.0**	12.8	**49.2**	**77.1**	**86.6**	**2.0**	8.2	**425.7**

Table 1. Performance on MSRVTT, **boldface** is the best results and underline is the runner-ups. Note that we have resolved the data-leakage problem in the original TMASS codes, and re-implemented XPool†, CLIP-VIP†, TMASS†, TV-ProxyNet†, and Video-ColBERT† using the CLIP4Clip codebase to align with other methods. * denotes the reproduced results, all of which are fine-tuned on ViT-B/32 or ViT-B/16.

where $\mathbb{I}_{ij}$ is a indicator for positive pairs $(+1)$ and negative pairs (-1), $\tau = e^{\tau_p}$, τ_p and b are learnable and initialized as $\tau_p = 4.77$ and $b = -12.93$, according to the pre-training in [56]. Finally, we replace $\mathcal{L}_{\text{ce}}$ with $\mathcal{L}_{\text{sig}}$ in Eq. (17):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sig}}(\mathbf{t}^{\text{gen}}, \mathbf{v}) + \lambda_{\text{sup}}\mathcal{L}_{\text{sig}}(\mathbf{t}^{\text{sup}}, \mathbf{v}) + \lambda_{\text{eam}}\mathcal{L}_{\text{eam}}. \quad (19)$$

4. Experiment

4.1. Experimental Settings

Datasets and Metrics. We use four benchmark datasets: **MSRVTT** [52], **DiDeMo** [1], **MSVD** [4], and **VATEX** [45]. These datasets contain sufficient video clips, each of which has multiple captions. We use the standard TVR metrics [33]: Recall (%) at rank 1/5/10 (R@1/5/10), Median Rank (MdR), Mean Rank (MnR), and Rsum (R@1+R@5+R@10); and also the total test time (sec.). More details can be found in Appendix E.

Implementation. We initialized our model by CLIP pre-trained models (ViT-B/32 and ViT-B/16) [37], so hidden dimension $d = 512$. We uniformly sample $M = 12$ frames per video (64 for DiDeMo) and the max length of text is 32 (64 for DiDeMo). In FRL, we sample $S = 20$ text candidates. For MCMC sampling in EAM, step number $K = 20$, step size $\eta = 1$, and noise variance $\sigma^2 = 0.005$. For loss coefficient, we set $\lambda_{\text{sup}} = 0.8$ following [43] and $\lambda_{\text{eam}} = 1.0$ by default. We train on two A100 GPUs for 5 epochs with batch size $B = 64$, weight decay 0.2, dropout 0.3, learning rate 10^{-7} for CLIP modules and 10^{-4} for non-CLIP modules; and we evaluate on one A100 with batch size 8. More details can be found in Appendix E.

4.2. Comparison to State-of-the-art Models

We compare our proposed EagleNet with existing state-of-the-art (SOTA) models [17, 27, 32, 33, 38, 41, 43, 48, 51, 53, 57]. We show the performance in Tabs. 1 and 2, and the test time in Fig. 1. See Appendix F for further analysis and complexity details. Note that we have addressed the data-leakage problem in the original TMASS codes, and re-implemented XPool†, CLIP-VIP†, TMASS†, TV-ProxyNet†, and Video-ColBERT† using CLIP4Clip codebase to align with other models for fair comparisons.

Overall, EagleNet achieves the best text-to-video R@1 performance across datasets, outperforming the second-best by 2.00% on MSRVTT, 2.59% on DiDeMo, and 2.41% on

	Method	DiDeMo R@1↑	DiDeMo R@5↑	DiDeMo R@10↑	MSVD R@1↑	MSVD R@5↑	MSVD R@10↑	VATEX R@1↑	VATEX R@5↑	VATEX R@10↑	Rsum↑
	CE [31]	16.1	41.1	82.7	19.8	49.0	-	-	-	-	-
	HGR [5]	-	-	-	-	-	-	35.1	73.5	83.5	-
	ClipBERT [26]	20.4	48.0	60.8	-	-	-	-	-	-	-
	SupportSet [36]	-	-	-	28.4	60.0	-	45.9	82.4	-	-
	Frozen [2]	31.0	59.8	72.4	33.7	64.7	-	-	-	-	-
	BridgeFormer [16]	37.0	62.2	73.9	-	-	-	-	-	-	-
	TVMM [28]	36.5	64.9	75.4	-	-	-	-	-	-	-
ViT-B/32	CLIP4Clip* [33]	40.6	67.7	77.8	46.2	75.6	84.6	56.6	87.5	93.6	630.2
ViT-B/32	XPool* [17]	43.5	69.5	79.0	47.1	76.5	85.2	57.5	87.8	93.7	639.8
ViT-B/32	TS2Net* [32]	39.5	69.7	80.9	44.7	75.5	84.9	57.7	88.7	94.2	635.8
ViT-B/32	CLIP-VIP* [53]	44.7	72.3	81.3	46.5	76.2	85.2	**59.3**	89.0	94.3	648.8
ViT-B/32	ProST* [27]	45.8	74.3	**84.5**	44.3	75.2	84.8	57.3	88.1	94.2	648.5
ViT-B/32	TMASS* [43]	43.4	70.6	81.0	41.8	71.9	81.4	56.1	88.1	93.8	628.1
ViT-B/32	RAP [3]	42.6	70.4	79.6	44.9	73.7	83.1	-	-	-	-
ViT-B/32	SHE-Net [54]	45.6	**75.6**	83.6	47.6	76.8	85.5	-	-	-	-
ViT-B/32	GLSCL* [57]	45.5	75.4	84.3	-	-	-	-	-	-	-
ViT-B/32	TempMe* [41]	**47.2**	73.7	81.3	-	-	-	-	-	-	-
ViT-B/32	TV-ProxyNet* [51]	46.8	72.1	81.3	46.6	76.0	85.0	58.9	89.1	**94.4**	650.2
ViT-B/32	Video-ColBERT* [38]	40.2	68.6	79.3	43.9	73.9	83.5	57.1	87.4	93.1	627.0
ViT-B/32	**EagleNet (Ours)**	46.8	74.0	83.5	**47.9**	**77.7**	**85.9**	59.0	**89.2**	94.1	**658.1**
ViT-B/16	CLIP4Clip* [33]	45.1	71.5	81.4	47.6	78.0	86.4	60.9	90.6	95.4	656.9
ViT-B/16	XPool* [17]	44.1	73.5	83.4	49.5	79.6	87.9	62.0	91.0	95.6	666.6
ViT-B/16	TS2Net* [32]	43.1	72.3	81.5	47.0	78.0	86.6	62.1	91.2	95.8	657.6
ViT-B/16	CLIP-VIP* [53]	47.2	75.5	84.4	49.2	79.1	87.5	63.8	91.6	**96.0**	674.3
ViT-B/16	ProST* [27]	47.0	75.5	84.2	48.1	78.6	87.4	62.0	91.3	95.8	669.9
ViT-B/16	TMASS* [43]	42.1	73.1	82.3	44.9	75.2	84.2	60.8	90.5	95.3	648.4
ViT-B/16	GLSCL* [57]	49.1	76.5	85.7	-	-	-	-	-	-	-
ViT-B/16	TempMe* [41]	50.2	76.4	85.0	-	-	-	-	-	-	-
ViT-B/16	TV-ProxyNet* [51]	47.9	75.2	84.5	49.7	79.7	88.0	**64.0**	**91.7**	95.9	676.6
ViT-B/16	Video-ColBERT* [38]	44.5	71.7	82.1	47.7	78.3	86.6	61.1	90.2	95.3	657.5
ViT-B/16	**EagleNet (Ours)**	**51.5**	**78.7**	**86.8**	**50.9**	**80.7**	**88.3**	63.6	91.5	95.7	**687.7**

Table 2. Text-to-video results on DiDeMo, MSVD, and VATEX. Note that we have resolved the data-leakage problem in the original TMASS codes, and re-implemented XPool†, CLIP-VIP†, TMASS†, TV-ProxyNet†, and Video-ColBERT† using CLIP4Clip codebase to align with other methods. * denotes the reproduced results, all of which are fine-tuned on ViT-B/32 or ViT-B/16.



Figure 3. Visualization of EagleNet, TMASS, and TV-ProxyNet.

MSVD, respectively. And our model consistently reaches the highest Rsum results, demonstrating the superiority and universality of our proposed model.

Both TMASS [43] and TV-ProxyNet [51] focus solely on the interactions between text and frames/video, so they fail to exploit temporal information within the video, resulting in suboptimal performance (MSRVTT text-to-video R@1 48.5% for TMASS and 49.5% for TV-ProxyNet). From Fig. 3, they might retrieve incorrect videos with similar semantics but different contexts. In contrast, EagleNet also leverages the internal frame-to-frame interactions within the video when expanding text semantics. Modeling such frame-level relationships enables the enriched text embedding to capture both the global video semantics and the temporal contextual information, leading to SOTA performance across benchmarks. Visualization of not considering frame context (without FRL) is presented in Appendix F.5.2. Moreover, TMASS, based on XPool [17], designs a stochastic text modeling strategy to sample text embeddings with richer semantics. However, after addressing the data-leakage issue, its text-to-video R@1 result (48.5%) is worse than that of XPool (49.2%) on MSRVTT, underperforms the runner-up by 16.14% on DiDeMo, and achieves unsatisfactory performance on MSVD (44.9%) and VATEX (60.8%). These results suggest that stochastically sampling a single text embedding per text-video pair may introduce undesired noise, degrading the semantic quality of the final text representation. EagleNet instead extracts frame-relevant information from multiple text candidates, producing an accurate text embedding. We provide detailed visualization comparisons of different models in Appendix F.5.1.

4.3. Ablation Study

4.3.1. The Effect of Different Components

From Tab. 3, we first analyze the effect of different components (FRL and EAM). The base model (Row 1, R@1 48.5% on MSRVTT and 42.1% on DiDeMo) only uses the interactions between text and frames, yielding worse results than the baseline, XPool (49.2% on MSRVTT and 44.1% on DiDeMo). In contrast, FRL (Row 2) incorporates frame-frame relationships and thus captures the frame contextual information, improving the final performance (R@1 48.8%

	Components			MSRVTT			DiDeMo			Avg		
	FRL	EAM	$\mathcal{L}_{\text{sig}}$	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
				48.5	74.1	83.4	42.1	73.1	82.3	45.3	73.6	82.9
	✓			48.8	75.6	83.9	47.9	75.2	84.1	48.4	75.4	84.0
		✓		49.0	75.3	84.0	43.4	73.6	82.7	46.2	74.5	83.4
	✓	✓		50.5	76.5	84.2	49.2	76.6	85.3	49.9	76.6	84.8
			✓	47.8	75.9	84.4	43.9	75.6	83.6	45.9	75.8	84.0
	✓		✓	49.9	**77.3**	84.3	49.1	77.5	85.0	49.5	77.4	84.7
		✓	✓	48.6	75.3	83.7	43.8	75.6	83.8	46.2	75.5	83.8
EagleNet	✓	✓	✓	**51.0**	76.2	**85.6**	**51.5**	**78.7**	**86.8**	**51.3**	**77.5**	**86.2**

Table 3. Text-to-video results on different components.

Energy Function	MSRVTT			DiDeMo		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
`CosSim`	50.0	**76.9**	85.5	50.2	78.3	86.3
`Bilinear`	**51.0**	76.2	**85.6**	51.3	78.3	86.6
`MLP`	50.3	76.6	85.1	**51.5**	**78.7**	**86.8**

Table 4. Text-to-video on different energy functions.

Pooling Form	MSRVTT			DiDeMo		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
w/o EAM	49.9	**77.3**	84.3	49.1	77.5	85.0
$E_\theta(\mathbf{t}, \mathbf{v})$	48.8	76.3	84.6	47.8	77.1	85.2
`Maxpool`	50.4	76.5	85.5	51.3	**79.0**	86.6
`Minpool`	50.5	77.2	84.9	50.9	78.5	86.5
`Avgpool`	**51.0**	76.2	**85.6**	**51.5**	78.7	**86.8**

Table 5. Text-to-video on different pooling approach for E_θ.

on MSRVTT and 47.9% on DiDeMo). Moreover, as mentioned in Sec. 4.2, the stochastic text modeling of TMASS (Row 1) may bring undesired noise, thereby also lowering the performance compared to its baseline, XPool. FRL instead extracts frame-related information from multiple text candidates, resulting in an accurate enriched text embedding. Meanwhile, EAM models the distribution of real text-video pairs from a fine-grained perspective, enhancing the ability of the model to consider fine-grained matching relevance and consequently improving the performance (Row 3, 49.0% on average). Combining them, EAM can enhance FRL's performance (Row 4, 49.9% on average) in learning relationships by utilizing the fine-grained information in text-frame interactions.

Moreover, we observe an overall improvement when using $\mathcal{L}_{\text{sig}}$ over $\mathcal{L}_{\text{ce}}$ (Row 5 to 8). Since sigmoid loss models each text-video pair independently, it is unaffected by the batch size and negative samples, enabling more effective alignment and stable optimization. These features make it more suitable for TVR scenarios.

Detailed visualization comparison of different components can be found in Appendix F.5.2.

4.3.2. The Selection of Energy Function

We compare the performance of different energy functions described in Sec. 3.2.2 (Tab. 4). On the dataset MSRVTT, the `Bilinear` shows the best R@1 results (51.0%). On DiDeMo, the best result (51.5%) is given by `MLP`. This indicates that we need to select a suitable one for different applications.

Although `CosSim` is learning-free and shows comparable performance, such a non-parametric energy function limits the capability of dealing with complicated scenarios. The `Bilinear` can be regarded as `CosSim` with learnable parameters and is more flexible. `MLP` adaptively learn the connection between text and video and is no longer limited to a cosine similarity form, so it can discover more implicit information in some cases but has more leanable parameters. Overall, we choose `Bilinear` as our final implementation.

4.3.3. The Pooling Approach for Text-Frame Energy

Here we discuss the pooling approach for energy function, `Avgpool` ($\frac{1}{M}\sum_{i=1}^{M} E_\theta(\mathbf{t}, \mathbf{f}_i)$), `Maxpool` ($\max_i E_\theta(\mathbf{t}, \mathbf{f}_i)$), and `Minpool` ($\min_i E_\theta(\mathbf{t}, \mathbf{f}_i)$). Since the energy of a matched pair is low, `Minpool` selects the frame that is most relevant to the given text. On the contrary, `Maxpool` selects the most irrelevant frame, which can be seen as a lower bound of text-video relevance. Besides, we can also use the global video embedding to calculate the text-video energy ($E_\theta(\mathbf{t}, \mathbf{v})$).

From Tab. 5, `Avgpool` shows the best R@1 performance on both MSRVTT (51.0%) and DiDeMo (51.5%), suggesting that considering all the text-frame interactions brings the most benefit to the retrieval. Regarding the other two approaches, selecting the most relevant frame yields better results on MSRVTT (`Minpool`, 50.5%); but improving the lower bound to enhance the overall relevance is more suitable for DiDeMo (`Maxpool`, 51.3%).

Besides, fusing frames before calculating energy shows lower performance (48.8% and 47.8%) than not using EAM (49.9% and 49.1%). Therefore, learning from fine-grained text-frame interactions can better improve the model in identifying the actual text-video pair. The fused video embedding provides only a global view, missing fine-grained details.

5. Conclusions

In this paper, we present EagleNet, a novel text-video retrieval framework that generates accurate and context-aware enriched text embeddings by exploring fine-grained relationships among text and video frames. We further employ an energy-based model to capture detailed interactions, which facilitates learning of relationships and enhances matching performance from a fine-grained perspective. Moreover, we instead use sigmoid loss for more effective alignment and stable training. Extensive experiments across four datasets demonstrate the superior performance of EagleNet compared with other state-of-the-art methods.

Acknowledgements

This work was supported in part by the Shenzhen Science and Technology Program under Grant No. KQTD20221101093559018, Grant No. CJGJZD20240729141505007, and Grant No. SYSRD20250529113401002, by the National Natural Science Foundation of China (NSFC) under Grant No. 62441619 and Grant No. 62411540034.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5803–5812, 2017. 6, 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew 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*, pages 1728–1738, 2021. 2, 6, 7, 3

[3] Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. Rap: Efficient text-video retrieval with sparse-and-correlated adapter. In *Findings of the Association for Computational Linguistics*, pages 7160–7174, 2024. 2, 3, 6, 7

[4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, 2011. 6, 3

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 1, 2, 3, 7

[6] Yuhan Chen, Yihong Luo, Yifan Song, Pengwen Dai, Jing Tang, and Xiaochun Cao. Decoupled graph energy-based model for node out-of-distribution detection on heterophilic graphs. In *International Conference on Learning Representations*, 2025. 5

[7] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 2

[8] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 33:8765–8775, 2020. 2, 5

[9] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 1, 2, 6, 3

[10] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021. 1, 2

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32:3608–3618, 2019. 5, 2

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 3

[13] Sheng Fang, Shuhui Wang, Junbao Zhuo, Qingming Huang, Bin Ma, Xiaoming Wei, and Xiaolin Wei. Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 4789–4800, 2022. 1, 2, 3

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of European Conference on Computer Vision*, pages 214–229, 2020. 1, 2, 6, 3

[15] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 1

[16] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 1, 2, 6, 7

[17] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei 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*, pages 5006–5015, 2022. 2, 3, 6, 7, 1

[18] Laura Hanu, James Thewlis, Yuki M Asano, and Christian Rupprecht. Vtc: Improving video-text retrieval with user comments. In *Proceedings of European Conference on Computer Vision*, pages 616–633, 2022. 2

[19] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, and Weiping Wang. Multi-feature graph attention network for cross-modal video-text retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 135–143, 2021. 1, 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[21] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *Proceedings of European Conference on Computer Vision*, pages 444–461, 2022. 2

[22] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306, 2022. 3

[23] Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. Hierarchical cross-modal

graph consistency learning for video-text retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1114–1124, 2021. 1, 3

[24] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. Association for Computational Linguistics (ACL), 2020. 2, 5

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. 3

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing 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*, pages 7331–7341, 2021. 1, 2, 6, 7, 3

[27] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4100–4110, 2023. 2, 3, 6, 7

[28] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. *Advances in Neural Information Processing Systems*, 35: 38655–38666, 2022. 6, 7, 3

[29] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925, 2021. 1, 2, 6, 3

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 5

[31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1, 2, 6, 7, 3

[32] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Proceedings of European Conference on Computer Vision*, pages 319–335, 2022. 2, 3, 6, 7

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 3, 6, 7

[34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 638–647, 2022. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3

[36] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2021. 6, 7, 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 6

[38] Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M de Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19691–19701, 2025. 2, 3, 6, 7

[39] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607, 2018. 3

[40] Leqi Shen, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Temporal modeling with frozen vision–language foundation models for parameter-efficient text–video retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 3, 6

[41] Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Yongjun Bao, Guiguang Ding, et al. Tempme: Video temporal token merging for efficient text-video retrieval. In *International Conference on Learning Representations*, 2025. 2, 3, 6, 7

[42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 3

[43] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16551–16560, 2024. 2, 3, 4, 5, 6, 7, 1

[44] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. Learning coarse-to-fine graph neural networks for video-text retrieval. *IEEE Transactions on Multimedia*, 23:2386–2397, 2020. 1, 3

[45] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 6, 3

[46] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 1, 2, 6

[47] Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky. Multi-query video retrieval. In *Proceedings of*

European Conference on Computer Vision, pages 233–249, 2022. 2

[48] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827, 2023. 2, 3, 6

[49] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019. 1, 2

[50] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3650–3660, 2021. 2

[51] Jian Xiao, Zhenzhen Hu, Jia Li, and Richang Hong. Text proxy: Decomposing retrieval from a 1-to-n relationship into n 1-to-1 relationships for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8655–8663, 2025. 2, 3, 6, 7, 5

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 6, 3

[53] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *International Conference on Learning Representations*, 2023. 2, 3, 6, 7

[54] Xuzheng Yu, Chen Jiang, Xingning Dong, Tian Gan, Ming Yang, and Qingpei Guo. She-net: Syntax-hierarchy-enhanced text-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2, 3, 6, 7

[55] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of European Conference on Computer Vision*, pages 471–487, 2018. 1, 2

[56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2, 5, 6, 3

[57] Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, Yihang Duan, Xinyu Lyu, and Heng Tao Shen. Text-video retrieval with global-local semantic consistent learning. *IEEE Transactions on Image Processing*, 34:3463–3474, 2025. 2, 3, 6, 7

[58] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 970–981, 2022. 1, 2, 3

[59] Xiaohan Zou, Changqiao Wu, Lele Cheng, and Zhongyuan Wang. Tokenflow: Rethinking fine-grained cross-modal alignment in vision-language retrieval. *arXiv preprint arXiv:2209.13822*, 2022. 2