Hybrid-Learning Video Moment Retrieval across Multi-Domain Labels

Weitong Cai
weitong.cai@qmul.ac.uk
Jiabo Huang
jiabo.huang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

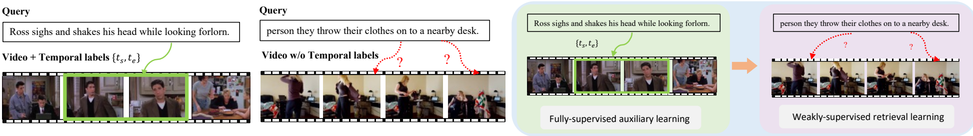
Computer Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London E1 4NS, UK

Abstract

Video moment retrieval (VMR) is to search for a visual temporal moment in an untrimmed raw video by a given text query description (sentence). Existing studies either start from collecting exhaustive frame-wise annotations on the temporal boundary of target moments (fully-supervised), or learn with only the video-level video-text pairing labels (weakly-supervised). The former is poor in generalisation to unknown concepts and/or novel scenes due to restricted dataset scale and diversity under expensive annotation costs; the latter is subject to visual-textual mis-correlations from incomplete labels. In this work, we introduce a new approach called *hybrid-learning video moment retrieval* to solve the problem by knowledge transfer through adapting the video-text matching relationships learned from a fully-supervised source domain to a weakly-labelled target domain when they do not share a common label space. Our aim is to explore shared universal knowledge between the two domains in order to improve model learning in the weakly-labelled target domain. Specifically, we introduce a *multiplE branch Video-text Alignment model* (EVA) that performs cross-modal (visual-textual) matching information sharing and multi-modal feature alignment to optimise domain-invariant visual and textual features as well as per-task discriminative joint video-text representations. Experiments show EVA's effectiveness in exploring temporal segment annotations in a source domain to help learn video moment retrieval without temporal labels in a target domain.

1 Introduction

Video moment retrieval (VMR) aims to locate a temporal moment in a long and untrimmed video by predicting its start and end time indices according to a natural language query sentence [13, 32]. This problem is intrinsically challenging as it requires not only to derive the semantic correspondences between video and text but also to recognise the subtle visual dissimilarity among different moments in the same unstructured videos of shared context. To learn effectively such subtle differences between similar video segments of different semantic descriptions, the exact temporal boundaries of target video moments are usually required in model training. However, such fine-grained temporal labelling is not only much more



(a) Fully-Supervised VMR (b) Weakly-Supervised VMR (c) Hybrid-Learning VMR

Figure 1: VMR tasks with different supervision settings.

time-consuming but also more ambiguous therefore subjective to more noise than still-image class annotations.

Existing methods take two approaches: **(1)** Fully-supervised methods [1, 7, 15, 19, 24, 25, 44] assume the availability of reliable and precise temporal boundary annotations for model training (Fig. 1(a)). However, considering the unaffordable annotation cost, the publicly available datasets are restricted in both scale and diversity. This results in poor generalisation of existing fully-supervised methods to new learning tasks in other domains [2, 45]. **(2)** Alternatively, weakly-supervised approaches [18, 22, 31, 32, 36] aims to reduce the annotation cost by model training with only video-level video-text pairing labels (Fig. 1(b)). However, it is harder for weakly-supervised learning to effectively derive semantically plausible video-text correspondences without temporal boundary labelling.

We consider that semantically plausible video-text correspondences should be agnostic to domains and can be shared across learning tasks. In this work, we formulate a new approach to solving this problem called *hybrid-learning video moment retrieval*, for performing jointly fully- and weakly-supervised learning using both a fully-labelled source domain and a weakly-labelled target domain simultaneously (Fig. 1(c)). This hybrid learning approach aims to optimise weakly-supervised learning of visual-textual correlations in a target domain without temporal labels of video moments by sharing knowledge on video-text alignment learned from a source domain. Such a problem is fundamentally challenging due to the distribution shift and vocabulary discrepancy across domains/learning tasks. Existing datasets are restricted to specific video scenes that are distinct between datasets, *e.g.*, ActivityNet-Captions (Anet) [20] involves mostly out-door activities whilst Charades-STA (Charades) [15] is mostly daily indoor routines. Videos in different domains have different data characteristics regarding visual and motion patterns, and video durations. Moreover, since the natural language queries are provided without explicit constraint on wording, existing datasets are different in their vocabulary, resulting in learning retrieval tasks in different label spaces, *e.g.*, TVR [21] uses *cat*, *kitty* and *tabby* to describe the concept of "cat" while *tabby* is missing in Anet and only *cat* is used in Charades. There are 41% and 61% of the words used in TVR and Anet training splits are shared cross domains. It is nontrivial to optimise simultaneously visual-textual correlations between both common vocabularies shared across tasks and similar/related semantics expressed in different words between domains.

In this work, we introduce a *multiplE branch Video-text Alignment model* (EVA) for hybrid-learning video moment retrieval across tasks in different domains. The key idea is to share the precise video-text temporal label information in fully-supervised learning in one domain with weakly-supervised learning in another domain of different data without temporal annotations. To that end, we formulate two concurrent learning branches with one weakly-supervised retrieval branch learning from target domain data and another fully-supervised auxiliary branch learning from a fully-labelled auxiliary dataset. For cross-domain cross-modal label sharing in model learning, we apply a cross-modal attention module to estimate the correlation between video and query and share this fine-grained matching knowledge to the weakly-supervised retrieval branch. Given the distribution shift and label inconsistency across domains (between datasets), we introduce a modality feature alignment constraint to mitigate feature discrepancies between different data sources in each single modality. This constraint is imposed by measuring the maximum mean discrepancy [17, 42] and applied to both before and after the cross-modal attention module. By doing so, the feature spaces of both video-text modalities are encouraged to share more areas in learning cross-modal attention to focus on video-text interaction between domains. Moreover, we further deploy a joint-modal domain classifier, combined with the weakly- and fully-supervised video moment retrieval losses for each retrieval task under different supervision in the two branches. This is to ensure that the joint video-text representations are per-task discriminatively optimised.

Our **contributions** are: **(1)** To our best knowledge, we make the first attempt at hybrid-learning VMR to explore jointly fully-labelled and weakly-labelled domains for optimising model learning of visual-textual correlations in the weakly-labelled domain where temporal labels are not given. **(2)** We introduce a multi-branch multi-modal formulation to transfer temporal label information as knowledge in model learning across tasks in different domains by both modal-specific and joint-modal feature alignments. **(3)** Extensive experiments show that the video-text alignment knowledge derived from the temporal labels in a source domain bring nonnegligible improvements to learn activity boundaries in a target domain without manual labels, ensuring EVA's competitiveness against the state-of-the-art VMR methods.

2 Related Works

Fully-supervised video moment retrieval. Video moment retrieval was first introduced in [1, 15] to derive the video-text alignment with the help of precise temporal boundary labels corresponding to natural language descriptions of video activities. To that end, MCN [1] generated candidate segments with different lengths in various starting points for retrieving. The studies of this problem mostly focus on either the multi-modal features encoding and fusion [6, 7, 10, 19, 24, 25, 44], or on generating high-quality candidate proposals [5, 23, 41, 46, 47, 48]. Although fully-supervised VMR methods have been shown effective on learning activity's boundary, they rely heavily on fine-grained manual temporal labels, which are not only labour-intensive and time-consuming but also ambiguity-prone and sensitive to subjective bias [1].

Weakly-supervised video moment retrieval. Different from the fully-supervised methods, there is no temporal boundary annotation being available for learning weakly-supervised approaches, liberated from the demand for large-scale temporal labels. Most existing weakly-supervised methods [31, 32, 37, 50] were based on contrastive learning, which is minimising the distance in video level between the videos and their matching queries from the dataset and maximising that of videos and the queries of others. Besides, some approaches [13, 22] tried to locate video moments by incorporating moment retrieval with sentence reconstruction. Recently, CRM [18] proposed to mine the cross-sentence relationship by exploiting the video-level paragraph descriptions as a whole. Regardless of their remarkable progress in the weakly-supervised setting, all of these approaches suffer from lacking fine-grained temporal annotations and overpass valuable temporal boundary labels available in existing datasets, resulting in avoidable waste.

Domain adaptation. There are many works in domain adaptation, especially unsupervised

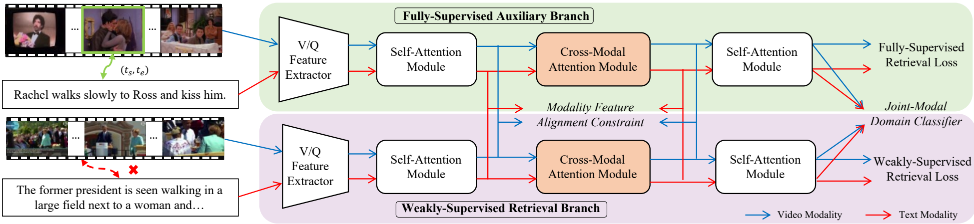


Figure 2: An overview of the proposed *multiplE branch Video-text Alignment model* (EVA) for hybrid-learning video moment retrieval.

domain adaptation, including variants of maximum mean discrepancies [17, 28, 30], adversarial learning [3, 26, 29] and transportable plan modelling [8, 11, 12] to measure and reduce the domain discrepancy. However, currently, there are a few works focusing on deep domain adaptation in multi-modal applications. Qi *et al.* [34] proposed a unified framework for the general multi-modal task with an adaptive modal fusion and domain constraints. Xu *et al.* [42] did domain adaptation on the video question answering task under a supervised domain adaptation setting. Chen *et al.* [9] and Liu *et al.* [27] proposed unsupervised domain adaptation methods for text-video retrieval from several trimmed video candidates. However, existing studies are mostly on knowledge adaptation across data domains while our objective is to adapt common video-text relationships across both domains and tasks with different vocabularies. Such a problem is both non-trivial and not straightforward to directly deploy existing methods.

3 Method

We construct a multiple branch network for model learning from a hybrid mixture labelled training data in two different domains (Fig. 2). Suppose there is an untrimmed video V with n_c consecutive disjoint clips $V = \{c_i\}_{i=1}^{n_c}$ in inflexible durations, where c_i is the i-th clip, and a corresponding query Q with n_w words $Q = \{w_i\}_{i=1}^{n_w}$, where w_i is the i-th word in the query sentence. In the *weakly-supervised retrieval branch*, videos V^w and corresponding queries Q^w from the temporally-unlabelled (target) dataset D^w are given. In model training, the *fully-supervised auxiliary branch* has the access to video-query pairs (V^f, Q^f) with temporal labels (t_s^f, t_e^f) from a fully-labelled auxiliary (source) dataset D^f. Our hybrid-learning VMR's goal is to locate the most likely time boundaries $(\hat{t}_s^w, \hat{t}_e^w)$ of the matching moment $\hat{S}^w$ to a given query Q^w by learning to optimise concurrently the pairing of both (V^w, Q^w) and (V^f, Q^f, t_s^f, t_e^f). Temporal labels (t_s^f, t_e^f) are mapped to their corresponding indices (i_s^f, i_e^f) in video clips V^f. For clarity concern, we use $*^w$ and $*^f$ to distinguish symbols from the weakly-supervised retrieval branch (target domain) and fully-supervised auxiliary branch (source domain) when needed and omit them in formulations/modules applied in both domains. For feature extraction, we utilise a pre-trained 3D-CNN model to acquire the clip-level video features with dimension d_c, and GloVe word embeddings [33] as word-level query features with dimension d_w. Both features are then projected into d-dimensional spaces by two independent fully-connected layers.

3.1 Multi-Branch Hybrid-Learning

The multi-branch hybrid-learning network is designed to augment a *weakly-supervised retrieval branch* on target domain unlabelled data by an additional *fully-supervised auxiliary branch* given a labelled auxiliary dataset.

Weakly-supervised retrieval branch. By exploring Transformer [38], EVA employs several attention modules to establish within- or cross-modality connections. Given $Y \in \mathbb{R}^{l_y \times d}$ and $X \in \mathbb{R}^{l_x \times d}$, the attention unit $Att(Y,X)$ attends Y using X as follows:

$$\mathcal{R}(Y,X) = \text{softmax}(YW^{q\top}W^kX^\top/\sqrt{d}), \quad Att(Y,X) = \text{FC}(Y + \mathcal{R}(Y,X)XW^{v\top}), \tag{1}$$

where the softmax$(\cdot)$ is the softmax normalisation by each row in the given matrix, and $\{W^q;W^k;W^v\} \in \mathbb{R}^{3\times d\times d}$ are three trainable weight vectors. A fully-connected layer FC$(\cdot)$ has the same dimension after projection. Thus, the within-modal self-attention modules for video and text modalities are:

$$V \leftarrow Att^V(V,V), \quad Q \leftarrow Att^Q(Q,Q), \tag{2}$$

which focus on exploring the within-modal dependencies by learning the correlations between pairs of elements in a video or a text sentence. We employ two sets of self-attention modules for video and text sequentially. Between them, the video is divided from clips into proposals $V = \{p_k\}_{k=1}^{n_p}$ with a sliding window strategy [22, 31]. Then we fuse the query Q and proposal feature p_k to construct a joint video-text representation j_k [15] and acquire matching score $P_m(p_k|Q)$ by a fully connected layer FC$(\cdot)$ and a sigmoid function $\sigma(\cdot)$:

$$P_m(p_k|Q) = \sigma(\text{FC}(j_k)), \quad j_k = (p_k + max(Q))\|(p_k \otimes max(Q))\|\text{FC}(p_k\|max(Q)), \tag{3}$$

where $+$, $\otimes$, $\|$ are the element-wise matrix addition, multiplication, and concatenation. Symbol $max(\cdot)$ denotes max-pooling to generate sentence-level query features by aggregating all the words in a query. The video-level matching score $P_m(V|Q)$ is obtained by the max-pooling of the proposal-level video-text matching score matrix $\{P_m(p_k|Q)\}_{k=1}^{n_p}$. For each positive pair (V,Q), two negative counterparts (V^-,Q) and (V,Q^-) are sampled by replacing either V or Q with V^- or Q^- from the mini-batch following [18]. A binary cross-entropy (BCE) loss is then used for weakly-supervised learning:

$$\mathcal{L}_w = 2*(-\log P_m(V|Q)) - \log(1-P_m(V|Q^-)) - \log(1-P_m(V^-|Q)). \tag{4}$$

Fully-supervised auxiliary branch. In this branch, we want to proactively explore existing temporally labelled video data from elsewhere to augment the weakly-supervised learning in the target domain. To project features to spaces with the same dimension as those in the weakly-supervised branch, and to boost the intra-modal contextual interaction in the source domain auxiliary dataset, we use the same attention unit design as in the weakly-supervised retrieval branch. Per Eq. (1), a cross-modal attention module is deployed to learn the video-text interaction by focusing on the most matched parts between the two modality spaces:

$$V \leftarrow Att^{Q\to V}(V,Q), \quad Q \leftarrow Att^{V\to Q}(Q,V), \tag{5}$$

and shared parameters to weakly-supervised branch, so as to promote more accurate cross-modal interaction in weakly-supervised retrieval learning. Then we compute video-query similarity score $S_{v-q} = V^f Q_m^f$, where Q_m^f is the modularized query vector from Q [21, 43], and use two 1D convolution filters to predict the start and end boundary in the score curves, computed by $P_{s/e} = \sigma(\text{conv1D}_{s/e}(S_{v-q}))$, and P_s, P_e denote the start and end probabilities.

The fully-supervised retrieval loss is a weighted combination of a retrieval loss $\mathcal{L}_r^f$ and the video-level BCE loss per Eq. (4):

$$\mathcal{L}_f = \lambda_r \mathcal{L}_r^f + \mathcal{L}_{bce}^f, \quad \mathcal{L}_r^f = H(P_s, I_s^f) + H(P_e, I_e^f), \tag{6}$$

where H is the cross-entropy function, (I_s^f, I_e^f) are the one-hot ground-truth labels for start and end indices (i_s^f, i_e^f), and λ_r is a trade-off hyper-parameter.

3.2 Multi-Modal Feature Alignment across Tasks and Domains

As discussed earlier, there are apparent domain distribution shifts and data characteristic differences among heterogeneous datasets in both video and text modalities, which bring challenges when utilising several datasets simultaneously. Here our goal is to bridge the gaps across tasks and domains, and to transfer the video-text matching relationships learned from the fully-supervised source domain to the weakly-labelled target domain effectively.

3.2.1 Modality Feature Alignment Constraint

Considering the differences in video and text modalities, diverse datasets may represent the same semantic information in different styles. The apparent domain gaps in both natural language and videos among datasets make any cross-modal interaction likely to be subject to domain biases. This will introduce noise to the moment-level temporal labels given by the external fully-labelled dataset when shared with the weakly-supervised learning process in a different domain. For more accurate cross-modal matching and more effective precise label information sharing, we propose to align the features both *before* and *after* the cross-modal attention module in both modalities respectively. Specifically, by quantising and minimising the distance of the distributions between the source and target domains, we constrain the learning of the source and target domain to let the feature spaces for video and text modalities share more common areas between the two domains. After aligning the modality spaces on both input and output sides of the shared cross-modal attention module, the video-text correlations learned on the fully-labelled auxiliary dataset is more generalisable and comprehensible to benefit the weakly-supervised retrieval. Motivated by [42], we use the maximum mean discrepancy (MMD) [17] to measure the distribution's distance and constrain the intermediate features' distributions between source and target domain in both video and text modalities by minimising it. Given source domain samples $D^s = \{\boldsymbol{x}_i\}_{i=1}^{n_s} \in \mathbb{R}^{n_s \times d}$ and target domain samples $D^t = \{\boldsymbol{y}_j\}_{j=1}^{n_t} \in \mathbb{R}^{n_t \times d}$, the MMD is calculated as :

$$M(D^s, D^t)^2 = \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} K(\boldsymbol{y}_i, \boldsymbol{y}_j) - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} K(\boldsymbol{x}_i, \boldsymbol{y}_j), \tag{7}$$

where K is the radial basis function kernel [35]. To align the modality features in both branches, we calculate the MMD loss both before $((V_b^f, V_b^w), (Q_b^f, Q_b^w))$ and after $((V_a^f, V_a^w),$ $(Q_a^f, Q_a^w))$ the cross-modal attention module between the fully-supervised auxiliary branch and the weakly-supervised retrieval branch in visual and text modalities respectively:

$$\mathcal{L}_{align} = (\lambda_{vid} M(V_b^f, V_b^w)^2 + M(Q_b^f, Q_b^w)^2) + (\lambda_{vid} M(V_a^f, V_a^w)^2 + M(Q_a^f, Q_a^w)^2), \tag{8}$$

where λ_{vid} in Eq. (8) is a trade-off hyper-parameter between video and query modalities.

3.2.2 Joint-Modal Domain Classifier

Given that different modality features have their own discriminative characteristics, we explore an adversarial strategy to bridge the domain gaps whilst keeping per-task discriminativeness, with the constraints of retrieval losses ($\mathcal{L}_w, \mathcal{L}_f$) in both branches. Here, we use a joint-modal domain classifier with the gradient reversal layer (GRL) [14, 34] to tackle the problem. The joint-modal domain classifier is designed to map the video feature V and query feature Q to a scalar domain label $label_d \in \{0,1\}$. The domain label is designed to distinguish whether the network input video and query are from the dataset with temporal boundaries (source domain) or the dataset without temporal boundaries (target domain). The mapping can be expressed as $label_d = G_d(\cdot)$. Firstly, we concatenate the video feature V and query feature Q after max-pooling to generate a video-text joint feature J. And we apply another MMD constraint on the V and Q. Then J is fed into the domain classifier $G_d(\cdot)$ which contains two fully-connected and a softmax layers:

$$G_d(J) = \mathrm{softmax}(\mathrm{FC}_1(\mathrm{FC}_2(J))). \tag{9}$$

For the purpose of optimising the model to get the features V and Q domain-invariant, we take the gradient reversal layer before the domain classifier $G_d(\cdot)$. The binary cross-entropy loss function is then adopted as the joint-modal domain classifier loss:

$$\mathcal{L}_{domain} = -\log(1 - G_d(J^f)) - \log G_d(J^w), \tag{10}$$

where J^f, J^w denote the joint features in the two branches respectively.

3.3 Model Training and Testing

In each training iteration, we randomly sample n videos with a pair of queries from the target temporal-unlabelled dataset and the same amount samples with time annotations from the external temporal-labelled dataset, as a mini-batch. The overall loss is computed by:

$$\mathcal{L} = \mathcal{L}_w + \lambda_f \mathcal{L}_f + \lambda_{align} \mathcal{L}_{align} - \lambda_{domain} \mathcal{L}_{domain}, \tag{11}$$

where $\lambda_f, \lambda_{align}$ and λ_{domain} are hyper-parameters for each loss. In test, only the weakly-supervised retrieval branch is deployed.

4 Experimental Results

Datasets. In experiments, we employed three commonly used VMR datasets: ActivityNet-Captions [4, 20], Charades-STA [15] and newly released TVR [21]. The statistics of them are shown in Table 1.

Dataset	#video	#moment			avg. len. (sec)		avg. len. (wrd)
		train	val	test	video	moment	query
Anet [4]	19290	37417	17505/17031	-	117.6	36.2	14.8
Charades [15]	6672	12408	-	3720	30.6	8.1	7.2
TVR [21]	21793	87175	10895	5445	76.2	9.1	13.4

Table 1: Statistics of VMR datasets.

For more extensive comparisons and to align with other weakly-supervised methods, we use Charades and Anet as the target datasets for weakly-supervised retrieval learning and comparative evaluation. For hybrid learning, we use TVR for the auxiliary training dataset with full temporal labelling, considering its large number of samples to cover greater linguistic diversity and precise video-text information, which is shown in Table 1.

Method	Source	Target	Charades			Anet		
			IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.1	IoU=0.3	IoU=0.5
2D-TAN [48]	✓	✗	14.65	4.30	1.26	40.16	28.71	17.29
XML [21]	✓	✗	32.49	18.27	8.87	31.78	17.18	9.27
MMN [39]	✓	✗	11.45	3.06	0.86	42.19	24.57	13.09
EVA	✓	✓	**62.01**	**40.21**	**18.22**	**74.09**	**49.89**	**29.43**

Table 2: Performance comparisons between EVA hybrid-learning and state-of-the-art fully-supervised VMR methods tested on Charades and Anet.

Method	Source	Target	IoU=0.3	IoU=0.5	IoU=0.7
TGA [32]	✗	✓	29.68	17.04	6.93
SCN [22]	✗	✓	42.96	23.58	9.97
LoGAN [37]	✗	✓	51.67	34.68	14.54
BAR [40]	✗	✓	44.97	27.04	12.23
RTBPN [49]	✗	✓	60.04	32.36	13.24
VLANet [31]	✗	✓	45.24	31.83	14.17
CCL [50]	✗	✓	-	33.21	15.68
CRM [18]	✗	✓	53.66	34.76	16.37
EVA	✓	✓	**62.01**	**40.21**	**18.22**

(a) Evaluated on Charades

Method	Source	Target	Split	IoU=0.1	IoU=0.3	IoU=0.5
WS-DEC [13]	✗	✓	val_1	62.71	41.98	23.34
WSLLN [16]	✗	✓	val_1	75.4	42.8	22.7
BAR [40]	✗	✓	val_1	-	49.03	30.73
CRM [18]	✗	✓	val_1	**76.66**	**51.17**	**31.67**
EVA	✓	✓	val_1	70.79	46.23	28.00
SCN [22]	✗	✓	val_2	71.48	47.23	29.22
RTBPN [49]	✗	✓	val_2	73.73	49.77	29.63
CCL [50]	✗	✓	val_2	-	50.12	31.07
CRM [18]	✗	✓	val_2	**81.61**	**55.26**	**32.19**
EVA	✓	✓	val_2	74.09	49.89	29.43

(b) Evaluated on Anet

Table 3: Comparisons with state-of-the-art weakly-supervised VMR methods.

Evaluation protocol. Following prior works [13, 40], we use $IoU = m$, to calculate the percentage of the top predicted moment having Intersection over Union (IoU) larger than m.

Implementation details. We used C3D features after PCA (500-D) for per-frame representations in Anet, I3D (1024-D) for Charades, and either C3D or I3D for TVR depending on the features used on the other domain. GloVe embeddings [33] were used as the word-level feature representations (300-D). The hidden features' dimension d for both video clips and word representations were 256-D. For the weakly-supervised retrieval branch, the sliding windows stride was 8 and the window sizes were $\{8, 16, 32, 64, 128\}$ in Anet and $\{8, 12, 20, 32, 64\}$ in Charades. The model was trained 50 epochs by Adam optimiser with a batch size of 64 and learning rate of $1e-4$. The trade-off hyper-parameters were set as $\lambda_r^f = 0.1, \lambda_f = 1, \lambda_{vid} = 0.8, \lambda_{domain} = 0.01, \lambda_{align} = 1$.

4.1 Comparisons with the State-of-the-art

Comparisons with fully-supervised methods. The comparative evaluations on EVA in Table 2 are designed for a practical scenario which does not assume the target new data was drawn from the identical distributions as the one used to train the model. Fully-supervised models rely heavily on manual temporal labels and lack the design to utilise or finetune on other datasets when temporal boundary labels are not available. In this case, the fully-supervised methods were trained with the TVR dataset with full temporal annotations in model training, and evaluated on Charades and Anet val_2 respectively. The 'Source' and 'Target' columns indicate the access of temporal-labelled and unlabelled data in the training stage. Table 2 shows all fully-supervised methods trained on a specific dataset suffer a serious performance degradation when deployed to a new domain in test, demonstrating their poor generalisation abilities. Our proposed EVA outperforms all these methods in all metrics. Compared to EVA, all these fully-supervised methods are not designed for and cannot simultaneously learn jointly from cross-domain hybrid labelling information given by both fully labelled and weakly labelled data in different domains. Our new multi-branch hybrid learning model demonstrates compellingly its ability to utilise and exploit effectively cross-domain different training labels.

Comparisons with weakly-supervised approaches. Table 3 compares EVA with the state-of-the-art weakly-supervised models. All these methods have no access to temporal labels in training. It is evident that EVA performs well on Charades which shows the ability to introduce precise video-text interaction information from an external heterogeneous temporal-labelled dataset. In Anet, some powerful weakly-supervised methods will maintain better performance. Our analysis is that some of these methods are likely to have overfitted a training dataset. The video-query pairs from prevailing datasets suffer annotation biases [45], where for instance the moments will fall into several specific time locations in both train and test splits. Due to this, a model trained with such a dataset may only make retrieval by selecting a moment from the frequency statistics of the bias-based training split and still have satisfactory performance on the test set which shares similarly biased distribution.

To test our assumption on weakly-supervised model overfitting, we further carried out experiments on distribution changed splits for ActivityNet-CD and Charades-CD [45] and evaluated on the out-of-distribution (OOD) test splits. Table 4 shows the results under the discounted recall metric [45], comparing EVA with the best performing state-of-the-art weakly-supervised models WS-DEC and CRM. It is evident that EVA outperforms both methods, showing that EVA has great multi-modal understanding by exploiting and sharing the precise video-text interaction information from an external dataset of different labels in a different domain context.

Dataset	Method	Source	Target	IoU=0.3	IoU=0.5	IoU=0.7
Anet	WS-DEC [ ]	✗	✓	17.00	7.17	1.82
	CRM [ ]	✗	✓	22.77	10.31	-
	EVA	✓	✓	**23.11**	**11.29**	**4.32**
Charades	WS-DEC [ ]	✗	✓	35.86	23.67	8.27
	EVA	✓	✓	**47.83**	**31.71**	**12.76**

Table 4: EVA hybrid-learning VMR results on ActivityNet-CD and Charades-CD OOD splits.

4.2 Discussion and Analysis

Ablation study. We examined the effectiveness of each proposed component in Table 5. 'WR' and 'FA' are abbreviations of the weakly-supervised retrieval and fully-supervised auxiliary branches, whilst the 'Align' and 'Domain' mean the modality feature alignment constraint and joint-modal domain classifier. In the 'WR + fine-tune' row, we pre-trained the WR branch on the source dataset with temporal labels and fine-tuned it on the target dataset without labels, and in 'WR + unlabelled source', the WR branch was trained jointly without any temporal labels in both the source and target domain training data. Our WR branch was trained with a temporal-unlabelled target dataset only as of the baseline, and all other mentioned methods have the access to both target and source datasets. The results show that simply having an external temporal-labelled dataset gains limited improvement in some metrics but not all. Critically, the performance is limited due to the multi-modal domain gaps. The modality feature alignment constraint we introduced in each modality and the joint-modal domain classifier proposed in EVA to align the single-modal and joint-modal features respectively not only have their own benefits, but also when they are both adopted together by aligning the features both in each modality and cross-modality, the model performance benefited more.

Effect of module sharing. To promote precise cross-modal matching information interaction in two branches, we shared the parameters of the cross-modal attention modules and maintain the independence of self-attention modules to focusing on each domain intra-modal interaction. We investigated its effect by comparing the prediction recall of EVA constructed with different sharing parts on Anet val_2 split in Table 6, displaying its advantages in video-

Method	mIoU	IoU=0.1	IoU=0.3	IoU=0.5
WR	32.96	71.48	48.06	28.74
WR + fine-tune	33.14	72.59	48.86	28.21
WR + unlabelled source	33.01	72.18	48.18	28.01
WR + FA	33.11	71.62	48.46	28.56
WR + FA + Align	33.83	73.35	49.60	28.80
WR + FA + Domain	33.66	73.14	49.32	28.92
WR + FA + Align + Domain	**34.27**	**74.09**	**49.89**	**29.43**

Table 5: Component ablation study of EVA on TVR (source) and Anet (target).

Shared Module(s)	mIoU	IoU=0.1	IoU=0.3	IoU=0.5
No Sharing	33.85	73.40	49.73	29.21
Self1 + Self2	34.08	73.63	49.84	29.33
Self1 + Cross	34.00	73.92	49.21	28.91
Self2 + Cross	33.88	73.81	49.42	28.80
Self1 + Self2 + Cross	33.70	73.92	49.06	28.30
Cross	**34.27**	**74.09**	**49.89**	**29.43**

Table 6: Effects of module sharing.

Model	Train Dataset		(t,a)→c				(t,a)→a			
	Source	Target	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.1	IoU=0.3	IoU=0.5
WR	✗	✓	22.68	34.48	17.84	6.23	31.50	69.62	46.05	26.23
EVA	✓	✓	**24.18**	**36.91**	**22.53**	**8.79**	**34.04**	**73.76**	**49.80**	**28.75**
			(t,c)→c				(t,c)→a			
			mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.1	IoU=0.3	IoU=0.5
WR	✗	✓	39.74	61.72	38.60	16.63	17.94	45.04	24.47	12.35
EVA	✓	✓	**40.08**	**62.01**	**40.21**	**18.22**	**22.04**	**52.05**	**31.81**	**16.57**

Table 7: Model generalisation evaluation on EVA hybrid-learning that utilises different datasets in training and test.

text knowledge sharing. 'Self1', 'Self2', and 'Cross' refer to the first, second self-attention module, and cross-modal attention module respectively.

Generalisation to unseen data. To explore the generalisation ability of EVA, we try to use different datasets in training and evaluation stages. Specifically, we trained EVA on TVR, Anet and tested on Charades ((t,a)→c), and then trained on TVR, Charades and tested on Anet val_2 ((t,c)→a). All the videos are processed by I3D to extract the features. The results in Table 7 show that EVA not only has improvements on the target dataset, but also improves performance in all metrics over the baseline (WR) on a heterogeneous dataset of a different domain, which reveals that bringing an extra temporal-labelled dataset in our method can largely avoid the model from converging to some bias-based distributions and carry more useful cross-modal information for more precise moment localisation.

5 Conclusion

In this work, we introduced a new hybrid-learning approach to VMR by formulating a novel multiple branch video-text alignment framework. EVA explores fine-grained cross-modal matching interaction information with a shared cross-modal attention module between two branches. EVA also employs a modality feature alignment constraint and joint-modal domain classifier to align the features in individual and multiple modalities so to preserve their per-task discriminativeness. Experiments show the advantages of EVA over existing methods and demonstrate the effectiveness of our hybrid learning model in improving cross-domain weakly-supervised learning.

Acknowledgements

This work was supported by the China Scholarship Council, Vision Semantics Limited, and the Alan Turing Institute Turing Fellowship.

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 International Conference on Computer Vision*, pages 5803–5812, 2017.

[2] Peijun Bao and Yadong Mu. Learning sample importance for cross-scenario video temporal grounding. *arXiv preprint arXiv:2201.02848*, 2022.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[5] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. Adversarial video moment retrieval by jointly modeling ranking and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 898–906, 2020.

[6] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. *arXiv preprint arXiv:2109.06085*, 2021.

[7] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, 2018.

[8] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Reweighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7976–7985, 2018.

[9] Qingchao Chen, Yang Liu, and Samuel Albanie. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1072–1080, 2021.

[10] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453, 2021.

[11] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[12] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 447–463, 2018.

[13] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3062–3072. Curran Associates, Inc., 2018.

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017.

[16] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*, 2019.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[18] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, October 2021.

[19] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 217–225, 2019.

[20] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.

[21] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 447–463. Springer, 2020.

[22] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020.

[23] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[24] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Fevelopment in Information Retrieval*, pages 15–24, 2018.

[25] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 843–851, 2018.

[26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in Neural Information Processing Systems*, 29:469–477, 2016.

[27] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964, June 2021.

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning*, pages 2208–2217. PMLR, 2017.

[31] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 156–171. Springer, 2020.

[32] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019.

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[34] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 429–437, 2018.

[35] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel methods in computational biology*. MIT press, 2004.

[36] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020.

[37] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[39] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. *arXiv preprint arXiv:2109.04872*, 2021.

[40] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1283–1291, 2020.

[41] Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. Natural language video localization with learnable moment proposals. *arXiv preprint arXiv:2109.10678*, 2021.

[42] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. *arXiv preprint arXiv:1911.04058*, 2019.

[43] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[44] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019.

[45] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021.

[46] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2215–2224, 2021.

[47] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019.

[48] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020.

[49] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4098–4106, 2020.

[50] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33, 2020.