Coarse-to-Fine Recurrently Aligned Transformer with Balance Tokens for Video Moment Retrieval and Highlight Detection

Yi Pan[1,2], Yujia Zhang[1,*], Hui Chang[1], Shiying Sun[1], Feihu Zhou[3], Xiaoguang Zhao[1]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Department of Critical Care Medicine, The First Medical Centre, Chinese PLA General Hospital
{panyi2022,zhangyujia2014,hui.chang,sunshiying2013,xiaoguang.zhao}@ia.ac.cn, zhoufh301@126.com

***Abstract*—Video moment retrieval (MR) and highlight detection (HD) are two user-oriented video understanding tasks aimed at extracting query-dependent or highlighted moments to provide valuable content for users. While many recent works have proposed solutions for the joint task of MR and HD leveraging transformer architecture, we argue that existing approaches have not adequately aligned the video and text modalities using basic transformer encoders, and have overlooked the misalignment between irrelevant video clips and text queries. To address these issues, we introduce COREBA: a Coarse-to-Fine Recurrently Aligned Transformer with Balance Tokens. Firstly, we design a plug-and-play Coarse-to-Fine Cross-modal interaction (CFC) module, replacing the original transformer encoder to align the two modalities in a progressive manner. Secondly, we present a novel Recurrent Alignment Mechanism (RAM) to deeply align the modalities in a recurrent fashion. Thirdly, to mitigate the misalignment problem, we append text queries with learnable Balance Tokens to restrict the text information fused with irrelevant clips. Extensive experiments validate the effectiveness and superiority of our proposed method.**

***Index Terms*—Video moment retrieval, video highlight detection, multimodal alignment**

I. Introduction

A vast number of videos are uploaded to the internet for users' entertainment or learning purposes. Videos, composed of consecutive frames, inherently provide richer semantic information compared to static images. However, they also contain more redundant information, making it inconvenient for users to find the most relevant segments they are interested in. To address this problem, several user-oriented video understanding tasks have emerged, such as video thumbnail generation [1, 2], video moment retrieval [3, 4] and video highlight detection [5, 6]. Specifically, video moment retrieval focuses on extracting temporal segments from untrimmed videos specified by user queries, while video highlight detection aims to identify significant clips within a video by assigning saliency scores to each clip.

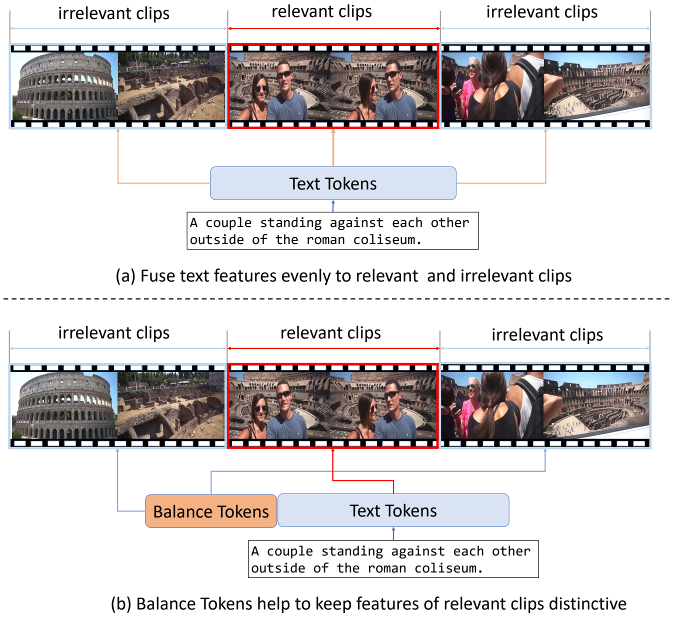

irrelevant clips
relevant clips
irrelevant clips
Text Tokens
A couple standing against each other outside of the roman coliseum.
(a) Fuse text features evenly to relevant and irrelevant clips
irrelevant clips
relevant clips
irrelevant clips
Balance Tokens
Text Tokens
A couple standing against each other outside of the roman coliseum.
(b) Balance Tokens help to keep features of relevant clips distinctive

Fig. 1: (a) The softmax operation in attention layers inevitably fuses text features into irrelevant video clips, resulting in the features of relevant clips becoming indistinct. (b) Incorporating learnable balance tokens helps confine text features fused into irrelevant clips, thereby preserving the distinctiveness of features associated with relevant clips.

In a recent study, Lei et al. [7] introduced a joint task of moment retrieval and highlight detection, along with the creation of the corresponding QVHighlights dataset. This task is particularly challenging as it requires a finer alignment between two vastly different modalities, namely, video and text, to accurately locate specified moments and assign correct saliency scores. The baseline model, Moment-DETR, is an end-to-end model based on the DETR [8] architecture and achieves promising accuracy. However, Moment-DETR aligns videos and text queries by simply concatenating tokens from both modalities, utilizing basic transformer encoders without employing refined alignment strategies. Many recent works, such as QD-DETR [9], MH-DETR [10], and UMT [11], delve deeper into the alignment challenge by introducing encoders equipped with multimodal cross-attention. They leverage video

*Corresponding author.

clip tokens as queries and text tokens as keys and values to derive text-dependent video representations. Nevertheless, we contend that directly fusing text features with video features poses difficulties due to the significant gap between the modalities. Furthermore, aligning all clips with the input sentence is impractical as the text query only specifies certain moments.

To address these limitations, we propose a **CO**arse-to-Fine **RE**currently Aligned Transformer with **BA**lance Tokens (COREBA). Firstly, we introduce a Coarse-to-Fine Cross-modal interaction (CFC) module for video-text alignment. This module initiates coarse alignment by fusing texts to videos through a VT-block and videos to texts through a TV-block. The roughly aligned tokens are then projected into a joint semantic space using a shared projection block before undergoing refined alignment through the same VT-block. Furthermore, we design a Recurrent Alignment Mechanism (RAM) that iteratively feeds features from two modalities into the same CFC module with multiple turns for deeper alignment. Experiments demonstrate this mechanism outperforms methods with multiple fusion layers. To address the misalignment issue between irrelevant video clips and texts, we introduce learnable Balance Tokens appended to original text tokens, as shown in Fig. 1. These tokens are trained to regulate the attention map, confining the injection of text information to irrelevant video clips and ensuring the distinctiveness of relevant clips. Comprehensive experiments validate the effectiveness of our proposed method, demonstrating its superiority over baselines on the QVHighlights dataset.

Our contributions can be summarized in three aspects: Firstly, we propose a plug-and-play Coarse-to-Fine Cross-modal (CFC) interaction module for effective video-text alignment. Secondly, we present a Recurrent Alignment Mechanism (RAM) to enable deeper multimodal fusion. Thirdly, we address the misalignment between irrelevant video clips and queries by introducing Balance Tokens. Comprehensive ablation experiments validate the effectiveness of each design.

II. Related Works

A. Video Moment Retrieval

The goal of video moment retrieval (MR) is to identify specific video segments given a text query. MR methods generally fall into two categories: proposal-based methods [3, 12] and proposal-free methods [13]. Proposal-based methods employ sliding windows [12] or anchors [3] to generate proposal candidates first, from which the required moments are selected. In contrast, proposal-free methods [13] directly predict the start and end of target moments in an end-to-end manner. Many recent studies [14, 15] have also adopted multi-modal transformer encoders to fuse video and text modalities and decoders to predict required moments. However, they use basic transformers without refined alignment strategies, potentially compromising the precision of moment localization.

B. Video Highlight Detection

Video highlight detection (HD) aims to identify highlight video clips by assigning a saliency score to each clip. The majority of methods, such as [16, 17], consider fully-supervised settings and rely on dense annotations at the frame level. Some recent approaches [5, 6] have introduced unsupervised methods that eliminate the need for dense annotations. These methods often utilize external concepts as weak supervision signals, including visual co-occurrence [18] or video duration [5].

C. Joint Video Moment Retrieval and Highlight Detection

Concurrently addressing moment retrieval and highlight detection in a unified manner is an emerging research topic leveraging both tasks' benefits. It enables precise moment localization, provides quick highlights, and shares common characteristics, as introduced by Lei et al. [7] in their baseline model, Moment-DETR, along with the QVHightlights dataset. Moment-DETR employs a standard transformer encoder-decoder architecture, coarsely aligning videos and texts by simply concatenating tokens before passing them through transformer encoder layers. Several recent works [9–11, 19, 20] have been proposed focusing on this challenging task based on Moment-DETR. UMT [11] utilizes a query generator to produce decoder queries, while QD-DETR [9] leverages a cross-modal encoder to extract query-dependent video features. EaTR [19] introduces an event reasoning module to generate event-aware moment queries. However, these approaches utilize basic transformer encoders to align two vastly different modalities, potentially limiting their ability to capture deeply fused multimodal representations. Additionally, they uniformly fuse text features to each video clip without considering the misalignment of irrelevant clips.

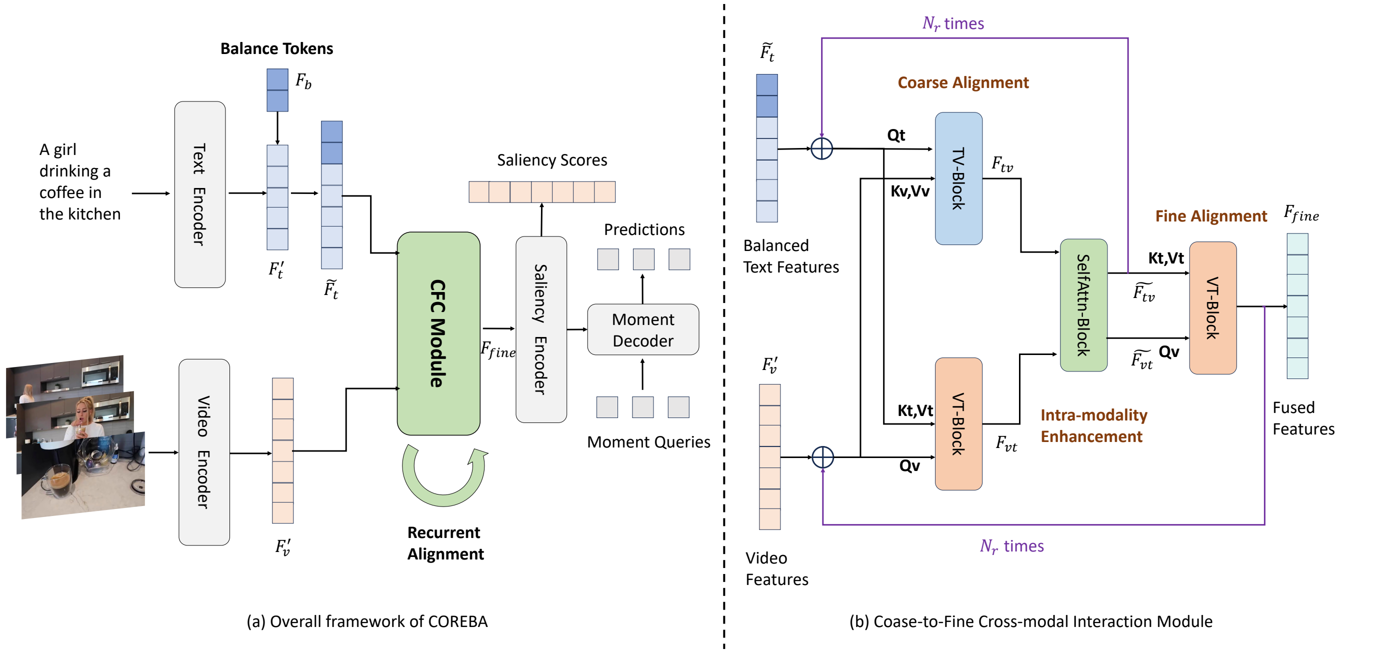
III. Approach

Here we first formalize the joint task of MR and HD, then proceed to describe our proposed method in detail. Given a video with L_v clips and a text query with L_t tokens, we utilize the extracted video features $F_v \in \mathbb{R}^{L_v \times d_v}$ and text features $F_t \in \mathbb{R}^{L_t \times d_t}$ provided in [7] (where d_v and d_t denote the feature dimensions of videos and texts respectively). The goal of the task is to locate N_m temporal segments represented by the start and end timestamps: $M = \{m_i \in \mathbb{R}^2\}_{i=1}^{N_m}$, and assign saliency scores $S \in \mathbb{R}^{L_v}$ to each clip of the video.

The overall architecture of our method is shown in Fig. 2(a). Initially, we augment text features with Balance Tokens and fuse the two modalities through the Coarse-to-Fine Cross-modal interaction module within the Recurrent Alignment Mechanism (RAM). Subsequently, leveraging the finely fused features $F_{fine} \in \mathbb{R}^{L_v \times d}$ (where d represents the dimension of the joint semantic space), we compute the saliency scores S using the saliency encoder and identify the required temporal segments M with a moment decoder.

A. Balance Tokens

We first project video features $F_v \in \mathbb{R}^{L_v \times d_v}$ and text features $F_t \in \mathbb{R}^{L_t \times d_t}$ to the same dimension d using two 3-layer MLPs denoted as MLP_1 and MLP_2: $F'_v = MLP_1(F_v) \in \mathbb{R}^{L_v \times d}$, $F'_t = MLP_2(F_t) \in \mathbb{R}^{L_t \times d}$, where F'_v



(a) Overall framework of COREBA

(b) Coase-to-Fine Cross-modal Interaction Module

Fig. 2: (a) The overall pipeline of our method. Balance tokens are appended to text tokens to restrict fusion with irrelevant video clips. Features from two modalities are passed to the Coarse-to-Fine Cross-Modal interaction (CFC) module within the Recurrent Alignment Mechanism for profound alignment. (b) The architecture of the CFC module. The two VT-blocks share the same weights (shown in the same color). We employ a self-attention block to align the two features in a shared semantic space. The fusion process repeats for N_r times. The residual operation is not illustrated for clarity.

and F'_t represent projected features. Previous works [9, 10, 19] typically fuse text features into entire video clips. However, the text query only specifies certain video segments. To balance attention weights and confine the fusion of text information with irrelevant video clips, we append learnable Balance Tokens to the original text tokens. Formally, let $F_b \in \mathbb{R}^{L_b \times d}$ denote L_b Balance Tokens. The balanced text feature sequence is given by:

$$\widetilde{F}_t = [F'_t || F_b] \in \mathbb{R}^{\widetilde{L}_t \times d} \tag{1}$$

where $||$ denotes the concatenation operation, and $\widetilde{L}_t = L_t + L_b$. The projected video features F'_v and balanced text features $\widetilde{F}_t$ are then passed to the Coarse-to-Fine Cross-Modal Interaction module for deeper fusion.

B. Coarse-to-Fine Cross-Modal Interaction Module

The Coarse-to-Fine Cross-modal interaction (CFC) module is introduced to achieve a progressive and in-depth alignment between video and text features. As depicted in Fig. 2(b), given the projected video features F'_v and the balanced text features $\widetilde{F}_t$, the module first fuse text features to each clip with a VT-block, which consists of a single cross-attention layer:

$$F_{vt} = CA_{vt}(F'_v, \widetilde{F}_t, \widetilde{F}_t) \in \mathbb{R}^{L_v \times d} \tag{2}$$

where CA_{vt} denotes the standard cross-attention operation with the video features F'_v as queries and the balanced text features $\widetilde{F}_t$ as keys and values. This operation introduces the textual information to video clips.

Similarly, we employ a TV-block to associate relevant text queries with video clips:

$$F_{tv} = CA_{tv}(\widetilde{F}_t, F'_v, F'_v) \in \mathbb{R}^{\widetilde{L}_t \times d} \tag{3}$$

where $\widetilde{F}_t$ serves as queries and F'_v serves as keys and values in the CA_{tv} cross-attention operation.

Here the fused features F_{vt} and F_{tv} demonstrate coarse alignment, as the cross-attention operation treats queries (video or text tokens) separately, disregarding intra-modality attributes such as the temporal relationships among video clips. Consequently, we utilize a self-attention block to further improve the fused representations by considering intra-modality attributes:

$$\begin{cases} F^s_{vt} = SA(F_{vt}, F_{vt}, F_{vt}) \\ \widetilde{F_{vt}} = F'_v + Mean(F^s_{vt}, F_{vt}) \\ F^s_{tv} = SA(F_{tv}, F_{tv}, F_{tv}) \\ \widetilde{F_{tv}} = \widetilde{F}_t + Mean(F^s_{tv}, F_{tv}) \end{cases} \tag{4}$$

Specifically, a shared self-attention block, denoted as SA is applied to features F_{vt} and F_{tv} to closely align them in a shared semantic space, producing F^s_{vt}, F^s_{tv}. The aligned features are individually subjected to mean averaging with F_{vt} and F_{tv}, and then fused features denoted as $\widetilde{F_{vt}}$ and $\widetilde{F_{tv}}$, are obtained through two residual paths. Finally, these two resulting features are fed into the same cross-attention block CA_{vt} for fine-grained alignment, yielding F_{fine}:

$$F_{fine} = CA_{vt}(\widetilde{F_{vt}}, \widetilde{F_{tv}}, \widetilde{F_{tv}}) \in \mathbb{R}^{L_v \times d} \tag{5}$$

Note that the Balance Tokens actively participate in the entire interaction process and are learnt automatically to balance the attention weights. The intricately aligned feature F_{fine} is further utilized by the Recurrent Alignment Mechanism before being passed to the saliency encoder and moment decoder.

C. Recurrent Alignment Mechanism

To further enhance the alignment between the video and text modalities, we design a Recurrent Alignment Mechanism that iteratively feeds features to the same CFC module for multiple turns. The pseudo code of this mechanism can be represented as follows:

Algorithm 1: Recurrent Alignment Mechanism

```
Input: $F'_v, \widetilde{F}_t$
Output: $F_{fine}$
for $i = 1, \ldots, N_r$ do
    $\widetilde{F}_{vt}, \widetilde{F}_{tv}, F_{fine} = CFC(F'_v, \widetilde{F}_t)$
    $F'_v = F'_v + F_{fine}$
    $\widetilde{F}_t = \widetilde{F}_t + \widetilde{F}_{tv}$
end
return $F_{fine}$
```

CFC denotes the operations in III-B, N_r is the number of recurrent turns, and $F'_v, \widetilde{F}_t, F_{fine}, \widetilde{F}_{vt}, \widetilde{F}_{tv}$ are features obtained from III-B. The module shares weights across the multiple alignment turns. In comparison to the direct utilization of multiple CFC modules, this mechanism enables the transformer layers to more effectively leverage features from various fusion stages. The experiments detailed in IV-E demonstrate the superiority of this mechanism.

D. Saliency Encoder and Moment Decoder

We adopt the design of the saliency encoder and moment decoder from QD-DETR [9]. Specifically, the feature F_{fine} is input to a two-layer transformer encoder ENC_s that includes a learnable saliency token $t_s \in \mathbb{R}^d$:

$$[\hat{F}_{fine}||\hat{t}_s] = ENC_s([F_{fine}||t_s]) \tag{6}$$

The saliency scores are computed by taking the products of the saliency token with all clip tokens:

$$S = \frac{\hat{F}^T_{fine} w_f \cdot \hat{t}_s w_s}{d} \in \mathbb{R}^{L_v} \tag{7}$$

where $w_f, w_s \in \mathbb{R}^d$ are learnable parameters.

For moment localization, we adopt the modified decoder from DAB-DETR [21] as described in [9]. Queries are designed by explicit centers m_c and lengths m_l of moments. With N_q queries, the decoder aggregates relevant clip features from $\hat{F}_{fine}$ and outputs features $F_{dec} \in \mathbb{R}^{N_q \times d}$.

We then apply a 3-layer MLP to the features to predict normalized moment centers and widths: $\hat{M} = \{\hat{m}_i\}_{i=1}^{N_q}, \hat{m}_i \in [0,1]^2$. Additionally, a linear layer is applied to predict moment labels: $\hat{Y} = \{\hat{y}_i\}_{i=1}^{N_q}$, where $\hat{y}_i \in \{0,1\}$. Here, 1 denotes *foreground* if the predicted moment matches a ground truth, and 0 denotes *background*. The ground truth moments and labels are denoted as $M = \{m_i\}_{i=1}^{N_q}$ and $Y = \{y_i\}_{i=1}^{N_q}$ with $\emptyset$ padded to match the number N_q. Following Moment-DETR [7], we employ the Hungarian algorithm to find an assignment between each (m_i, y_i) and the corresponding prediction $(\hat{m}_{\sigma(i)}, \hat{y}_{\sigma(i)})$.

E. Training Losses

We use the same losses as described in [9]. For moment retrieval, the loss between m_i and $\hat{m}_i$ comprises an $L1$ regression loss, a generalized IoU loss and a cross-entropy loss:

$$\begin{aligned} L_{mr} = \lambda_{L1}||m_i - \hat{m}_i|| \\ + \lambda_{gIoU} L_{gIoU}(m_i, \hat{m}_i) \\ - \lambda_{CE}[y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)] \end{aligned} \tag{8}$$

where λ_* are hyperparameters to balance different losses.

For highlight detection, we utilize three types of losses: the margin saliency loss between pairs of high-rank clips and low-rank clips, the negative saliency loss of unpaired videos and queries, and the rank-aware contrastive loss for contrastive learning:

$$\begin{cases} L_{marg} = max(0, \Delta + S(t^{low}) - S(t^{high})) \\ L_{neg} = -\log(1 - S(t^{neg})) \\ L_{cont} = -\Sigma_{r=1}^{R} \log \dfrac{\Sigma_{t \in T_r^{pos}} \exp\left(S(t)/\tau\right)}{\Sigma_{t \in (T_r^{pos} \cup T_r^{neg})} \exp\left(S(t)/\tau\right)} \end{cases} \tag{9}$$

where Δ is a positive margin for the margin loss, $S(\cdot)$ denotes the predicted saliency score of a given clip, t^{high} and t^{low} denote clips with high scores and low scores, respectively, t^{neg} denotes clips from unpaired videos, and r denotes the ranking variable from 1 to R (the maximum rank value). For further details, please refer to [9, 22].

The loss for highlight detection is then calculated as:

$$L_{hl} = \lambda_{marg} L_{marg} + \lambda_{neg} L_{neg} + \lambda_{cont} L_{cont} \tag{10}$$

where λ_{marg}, λ_{neg}, λ_{cont} are hyperparameters to balance the losses.

The overall training loss is a combination of the moment retrieval loss and the highlight detection loss:

$$L = L_{mr} + L_{hl} \tag{11}$$

IV. Experiments

A. Dataset

We conduct experiments on the recently proposed QVHighlights [7] dataset, which is currently the only dataset for the joint moment retrieval and highlight detection task. The dataset contains more than 10,000 videos annotated with free-form queries, one or more relevant moments and clip-wise saliency scores.

Method	Moment Retrieval					Highlight Detection	
	R1		mAP			>= Very Good	
	@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
Moment-DETR [7]	53.94	34.84	-	-	32.20	35.65	55.55
UMT [11]	60.26	44.26	56.70	39.90	38.59	39.85	64.19
MH-DETR [10]	60.84	44.90	60.76	39.64	39.26	38.77	61.74
UniVGT [20]	59.74	-	-	-	36.13	38.83	61.81
EaTR [19]	61.36	45.79	61.86	41.91	41.74	37.15	58.65
QD-DETR [9]	62.68	46.66	62.23	41.82	41.22	39.13	63.03
COREBA	**65.16**	**49.48**	**63.90**	**42.61**	**42.48**	**40.39**	**65.42**

TABLE I: Comparison with other methods shows that our method outperforms baselines across all metrics.

N_b	Moment Retrieval					Highlight Detection	
	R1		mAP			>= Very Good	
	@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
0	63.23	47.48	62.66	41.75	41.63	39.08	62.58
1	63.94 (+0.71)	47.23	63.17 (+0.51)	41.68	41.17	39.57 (+0.49)	63.29 (+0.71)
3	63.35 (+0.12)	48.26 (+0.78)	62.93 (+0.27)	42.59 (+0.84)	42.18 (+0.55)	39.55 (+0.47)	62.90 (+0.32)
5	62.45	46.32	61.97	41.54	40.99	39.33 (+0.47)	62.97 (+0.32)
7	**65.16** (+1.93)	**49.48** (+2.00)	**63.90** (+1.24)	**42.61** (+0.86)	**42.48** (+0.85)	**40.39** (+1.31)	**65.42** (+2.84)
9	62.68	46.66	62.23	41.82	41.22	39.13 (+0.05)	63.03 (+0.45)

TABLE II: Models with different numbers of Balance Tokens. N_b denotes the number of balance tokens.

id	N_{tv}	N_{vt}	N_s	Moment Retrieval					Highlight Detection	
				R1		mAP			>= Very Good	
				@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
(qd3)	0	3	0	62.00	46.52	61.84	41.68	41.53	39.40	63.81
(qd4)	0	4	0	63.94	48.71	63.33	42.57	42.10	40.04	65.10
(a)	0	1	1	63.42	48.00	63.47	42.63	42.29	39.93	64.58
(b)	1	1	0	62.84	46.26	63.08	42.17	42.05	39.79	62.77
(c)	1	2	1	63.68	47.68	62.14	41.86	41.43	39.85	64.00
(d)	1	1	2	64.32	48.52	62.94	**44.14**	**42.83**	39.79	63.16
(e)	1	1	1	**65.16**	**49.48**	**63.90**	42.61	42.48	**40.39**	**65.42**

TABLE III: Models with different TV-block, VT-block, Self-Attn Settings. N_{tv}, N_{vt} and N_s denote the number of TV-block, VT-block and Self-attention block, respectively.

B. Implementation Details

We use the official features provided in Moment-DETR [7]. Specifically, for videos we use the SlowFast [23] and CLIP [24] (ViT-B/32) features from the visual encoder, and for text queries, we use the CLIP features from the text encoder. The number of Balance Tokens is set to 7. The model consists of 1 CFC module, 2 saliency encoder layers and 2 decoder layers. The number of loops N_r in the Recurrent Alignment Mechanism is set to 3. The hidden dimension d of the model is 256. For losses, we follow the setting in QD-DETR [9] and set the parameters $\lambda_{marg} = 1, \lambda_{cont} = 1, \lambda_{L1} = 10, \lambda_{gIoU} = 1, \lambda_{CE} = 4, \lambda_{neg} = 1$. The model is trained for 200 epochs with a learning rate set to 1e-4 and batch size set to 32.

C. Evaluation Metrics

Following [7], we report the mean average precision (mAP) for the moment retrieval task. The mAP is computed with IoU thresholds of 0.5, 0.75 and an average over multiple IoU thresholds $[0.5 : 0.05 : 0.95]$. Additionally, we report the standard metric Recall@1 (R@1) used in single moment retrieval, where a prediction is considered positive if it has a high IoU with one of the ground truths. For highlight detection, we provide the mAP and the hit ratio [25] (HIT@1) for the highest scored clip.

D. Quantitative Results

We compare our method with recent state-of-the-art methods on QVHighlights *val* split. As shown in Table I, our method outperforms these methods across all metrics. For moment retrieval, we outperform the QD-DETR model by nearly 3% in R1 metric, and 0.74% to 1.67% in mAP metrics. For highlight detection, we outperform UMT by 1.23% in the HIT@1 metric. These results show the effectiveness and superiority of our method.

E. Analysis

1) Ablation on Balance Tokens: Firstly, we analyze the effectiveness of Balance Tokens with different numbers. We equip the model with $0, 1, 3, 5, 7, 9$ Balance Tokens respectively. The results are shown in Table II. It can be observed that different numbers of Balance Tokens boost the performance to varying degrees. Models with 7 Balance Tokens

Settings	Moment Retrieval					Highlight Detection	
	R1		mAP			>= Very Good	
	@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
MOD_2	63.16	47.74	**63.70**	42.79	**42.54**	39.70	**63.61**
REC_2	**64.71**	**47.94**	62.83	**42.86**	42.09	**39.98**	63.55
MOD_3	64.19	49.48	63.35	42.24	42.10	40.15	64.39
REC_3	**65.16**	**49.48**	**63.90**	**42.61**	**42.48**	**40.39**	**65.42**
MOD_4	63.68	47.61	**63.91**	**43.00**	**42.51**	39.37	62.90
REC_4	**64.77**	**47.74**	63.81	42.06	41.78	**39.77**	**65.42**
MOD_5	62.19	45.74	61.67	40.14	40.07	38.83	61.74
REC_5	**62.71**	**47.29**	**62.64**	**41.74**	**41.65**	**39.74**	**63.94**

TABLE IV: Comparison between Recurrent Alignment and Multiple Layers designs. REC_N refers to models with 1 CFC module and N recurrent loops. MOD_N refers to models with N CFC modules and no recurrent loops applied.

	RAM	CFC	BAT	MR					HD	
				R1		mAP			>= Very Good	
				@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
(a)				62.00	45.61	61.61	40.84	39.97	38.18	60.65
(b)	✓			62.65	47.81	62.72	**42.68**	42.26	39.01	62.77
(c)		✓		64.77	48.52	63.64	42.25	41.52	39.35	63.94
(d)			✓	61.35	47.10	61.51	41.90	41.45	39.11	63.29
(e)	✓	✓		63.23	47.78	62.66	41.75	41.63	39.08	62.58
(f)	✓		✓	63.16	48.00	63.33	42.43	42.20	39.29	63.42
(g)		✓	✓	65.03	48.45	63.56	42.41	42.38	39.79	64.06
(i)	✓	✓	✓	**65.16**	**49.48**	**63.90**	42.61	**42.48**	**40.39**	**65.42**

TABLE V: Ablation study on combinations of designs. BAT, CFC, and RAM refer to Balance Tokens, Coarse-to-Fine Cross-modal interaction module and the Recurrent Alignment Mechanism, respectively.

perform the best, achieving a 2% improvement in R1@0.7 and 2.84% improvement in HIT@1. Furthermore, in the highlight detection task, all models with Balance Tokens outperform the models without Balance Tokens. The results in the dense clip-level highlight detection task indicate that introduced Balance Tokens may effectively restrict the information from text features injected into irrelevant clips, benefiting the entire task at the clip-level.

2) Ablation on CFC module: Secondly, we analyze the design of the CFC module by exploring several alternative architectures. As detailed in III-B, our approach involves employing one TV-block, one shared VT-block twice ((2) and (5)), and one shared self-attention (4). The alternative architectures are designed as models without a TV-block, without a self-attention block, with two separate VT-blocks and with two separate self-attention blocks respectively. The experimental results in Table III illustrate that the model configured with the aforementioned default settings attains the highest overall performance.

The comparison between (a) and (e) reveals that it is beneficial to fuse video features into text features, which has been overlooked by previous studies [9, 10]. Experiments (b) and (e) showcase the value of intra-modality mining through a self-attention block, while (c) and (d) demonstrate the effectiveness of employing a shared VT-block and self-attention layer to align the features within a unified semantic space.

To prove that the performance is attributed to the architecture rather than the parameter count, we conduct additional experiments employing a model with 3 and 4 fusion layers introduced in QD-DETR, along with the same number of Balance Tokens (experiments (qd3) and (qd4)). The results demonstrate that our method surpasses QD-DETR with a similar parameter count.

3) Ablation on Recurrent Alignment Mechanism: Next, we conduct experiments on the Recurrent Alignment Mechanism (RAM). We specifically focus on whether the recurrent design holds superiority over employing multiple fusion modules. Thus, we conduct a comparative analysis between models with various numbers of recurrent alignment loops, and models with identical numbers of fusion modules but lacking recurrent alignment. As shown in Table IV, across most metrics, models incorporating recurrent alignment consistently outperform corresponding models employing multiple modules. This indicates that despite having fewer parameters, the recurrent alignment mechanism succeeds in aligning the features from different recurrent stages, enabling a more profound alignment.

4) Ablation on Combination of Designs: Finally, we conduct ablation studies concerning the combination of the three designs: Balance Tokens (BAT), the CFC module and the Recurrent Alignment Mechanism (RAM). Commencing with a QD-DETR baseline model with one cross-model fusion layer, we progressively incorporate our designs in various combinations. The results are summarized in Table V.

From experiments (b), (c), and (d), it is evident that incorporating one of the three designs contributes to performance enhancement. Among these designs, the CFC module demonstrates the most significant impact. Experiments (f) and (g) further illustrate that augmenting Balance Tokens to models

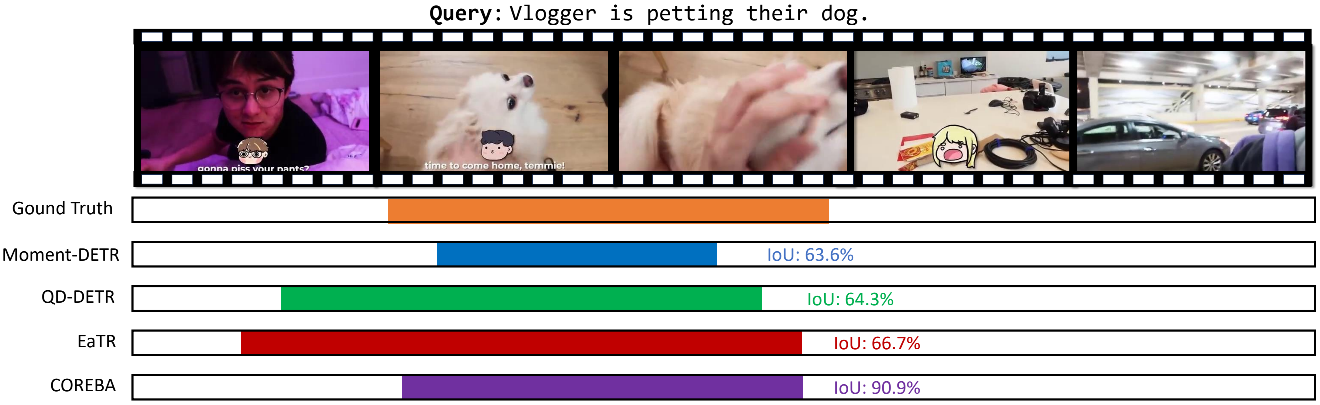


Fig. 3: Visualization of one prediction from COREBA and other methods. Our method locates the required moment more precisely.

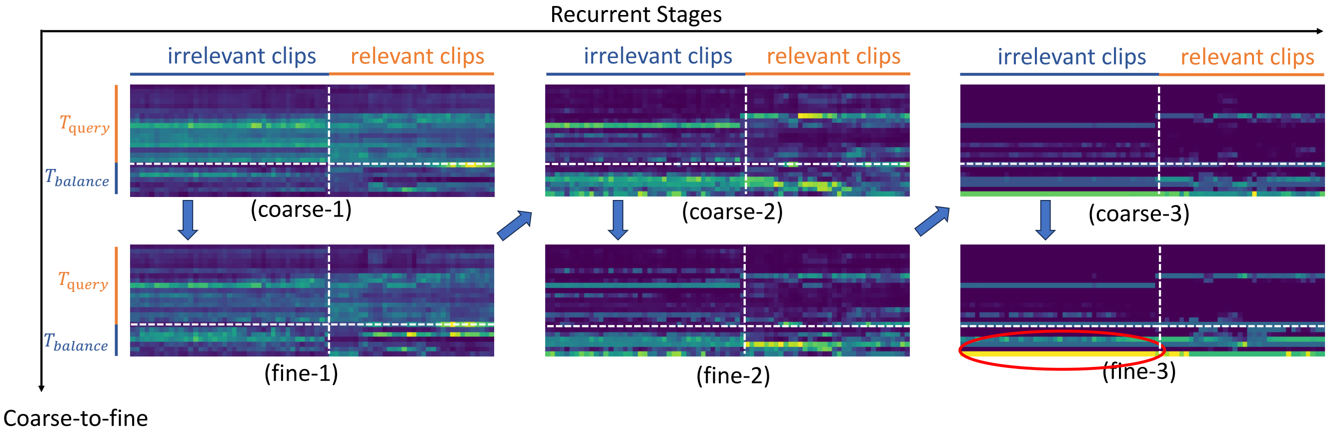


Fig. 4: Visualization of attention maps for the VT-block. Figure **coarse-i** or **fine-i** visualizes the attention map from the **coarse** or **fine** alignment in the **i-th** recurrent stage. T_{query} and $T_{balance}$ denote query tokens and balance tokens, respectively. The irrelevant clips and relevant clips are illustrated based on the ground truth. It is clear that the balance tokens gradually extract more attention from irrelevant clips during the alignment process, ensuring the distinctiveness of relevant clips (shown in the red circle).

with RAM or CFC improves performance. However, it is noteworthy that the model in experiment (e), combining RAM and CFC module, underperforms the model solely utilizing the CFC module. This outcome suggests that lacking Balance Tokens might lead to an excessive fusion of text features with irrelevant clips, resulting in decreased performance. Ultimately, the combination of all three designs yields the highest overall performance.

F. Visualization

We visualize some examples from the QVHighlights dataset. As illustrated in Fig. 3, our method demonstrates improved precision in locating specified moments. Additionally, we provide a visualization of the attention map from the VT-block of the CFC module, showcased in Fig. 4. During the coarse-to-fine recurrent alignment, the balance tokens progressively attract more attention from irrelevant clip tokens, thereby ensuring the distinctiveness of relevant segments.

V. Conclusion

In this paper, we propose COREBA, a Coarse-to-Fine Recurrently Aligned Transformer with Balance Tokens designed for the joint video moment retrieval and highlight detection task. We propose a plug-and-play Coarse-to-Fine Cross-model interaction module to effectively fuse videos and texts. A Recurrent Alignment Mechanism is further applied to the CFC module, enabling a profound alignment between the two modalities. Considering the misalignment between irrelevant video clips and text queries, we append learnable Balance Tokens to text queries, regulating the attention map and confining the fusion of text features with irrelevant clips. Extensive experiments demonstrate the superiority and effectiveness of our proposed method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62103410 and 62203438), the International Partnership Program of the Chinese Academy of Sciences (Grant No. 104GJHZ2023053FN), and the Independent Research Project of Medical Engineering Laboratory of Chinese PLA General Hospital (2022SYSZZKY12).

References

[1] Y. Yuan, L. Ma, and W. Zhu, “Sentence specified dynamic video thumbnail generation,” in *Proceedings of*

the 27th ACM International Conference on Multimedia, 2019, pp. 2332–2340.

[2] J. Wu, Y. Zhang, and X. Zhao, "Visual enhanced hierarchical network for sentence-based video thumbnail generation," *Applied Intelligence*, vol. 53, no. 19, pp. 22 565–22 581, 2023.

[3] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 870–12 877.

[4] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1–10.

[5] B. Xiong, Y. Kalantidis, D. Ghadiyaram, and K. Grauman, "Less is more: Learning highlight detection from video duration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1258–1267.

[6] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Contrastive learning for unsupervised video highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 042–14 052.

[7] J. Lei, T. L. Berg, and M. Bansal, "Detecting moments and highlights in videos via natural language queries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 846–11 858, 2021.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[9] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 023–23 033.

[10] Y. Xu, Y. Sun, Y. Li, Y. Shi, X. Zhu, and S. Du, "Mh-detr: Video moment and highlight detection with cross-modal transformer," *arXiv preprint arXiv:2305.00355*, 2023.

[11] Y. Liu, S. Li, Y. Wu, C.-W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3042–3051.

[12] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5267–5275.

[13] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6543–6554.

[14] C. Rodriguez-Opazo, E. Marrese-Taylor, B. Fernando, H. Takamura, and Q. Wu, "Locformer: Enabling transformers to perform temporal moment localization on long untrimmed videos with a feature sampling approach," *arXiv preprint arXiv:2112.10066*, 2021.

[15] W. Wang, J. Cheng, and S. Liu, "Dct-net: A deep co-interactive transformer network for video temporal grounding," *Image and Vision Computing*, vol. 110, p. 104183, 2021.

[16] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Joint visual and audio learning for video highlight detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8127–8137.

[17] M. Rochan, M. K. Krishna Reddy, L. Ye, and Y. Wang, "Adaptive video highlight detection by learning from user history," in *European Conference on Computer Vision*. Springer, 2020, pp. 261–278.

[18] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3584–3592.

[19] J. Jang, J. Park, J. Kim, H. Kwon, and K. Sohn, "Knowing where to focus: Event-aware transformer for video grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 846–13 856.

[20] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, "Univtg: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2794–2804.

[21] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," in *International Conference on Learning Representations*, 2021.

[22] D. T. Hoffmann, N. Behrmann, J. Gall, T. Brox, and M. Noroozi, "Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 897–905.

[23] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[25] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3707–3715.