TEXTUAL AND TEMPORAL-GUIDED FEATURE DECOUPLING FOR VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent approaches to Video Temporal Grounding (VTG) predominantly rely on CLIP-based representations, often augmented with visual encoders such as SlowFast or C3D to enhance temporal modeling. However, the prevailing "concat-then-project" paradigm disrupts the inherent alignment between CLIP's visual and textual modalities and undermines the temporal modeling capabilities of the additional video encoder. To address these, we propose FDAP, a plug-and-play Feature Decoupling and Aggregation Paradigm. FDAP introduces two key components: a Textual-Guided Feature Decoupling Module (TGFDM) that preserves CLIP's cross-modal alignment and SlowFast's temporal modeling via independent attention maps, and a Dual-branch Feature Aggregation Module (DFAM) that dynamically adjusts feature weights during aggregation based on query-specific needs. Extensive experiments across four VTG methods (M-DETR, TR-DETR, CG-DETR, Flash-VTG) on three benchmark datasets (QVHighlights, Charades-STA, TACoS) demonstrate consistent performance gains, *e.g.*, a 3% improvement in M-DETR's R1@0.7 metric. With minimal overhead (0.2M additional parameters), FDAP advances VTG feature modeling and generalizes effectively across diverse methods.

1 INTRODUCTION

With the rapid expansion of video content across diverse domains, efficiently identifying relevant segments within untrimmed videos has become a critical research challenge. The Video Temporal Grounding (VTG) task encompasses two downstream tasks: Moment Retrieval (MR) and Highlight Detection (HD), which aim to leverage both textual and temporal information to precisely localize video timestamps that correspond to the given textual descriptions.

The overall VTG task can be broadly divided into two key components: feature modeling for multi-modal alignment and inter-frame temporal information aggregation, and head design for task-specific adaptation to MR and HD. The majority of existing efforts have concentrated on head design Yan et al. (2023); Xiao et al. (2024); Zhang et al. (2020b;a); Cao et al. (2025). These methods generally leverage simple feature modeling approaches, which can be divided into two main frameworks: a) The "CLIP-Only" framework that relies solely on the multi-modal alignment capability of the pretrained CLIP Liu et al. (2024); Ju et al. (2022); Huang et al. (2023). b) The "CLIP+" framework that introduces an additional video encoder to integrate the temporal modeling capability Lin et al. (2023); Yang et al. (2024); Jiang et al. (2024); Zhao et al. (2024).

However, these feature modeling frameworks fail to meet the requirements of the VTG task. The "CLIP-Only" framework lacks temporal modeling capabilities. The "CLIP+" framework directly concatenates the features extracted from multiple encoders and projects them into a fixed dimension before being aligned with CLIP textual features Moon et al. (2023a;b); Sun et al. (2024); Liu et al. (2022); Jang et al. (2023). It disrupts both the multi-modal alignment of CLIP and the temporal modeling capabilities of the video encoder (see Figure 3). Moreover, in the VTG task, different queries typically have different preferences for textual and temporal features (see Figure 4), yet existing methods fail to model these preferences. Despite these shortcomings, feature modeling has long been overlooked in the research community.

To address these challenges, we propose a plug-and-play Feature Decoupling and Aggregation Paradigm (FDAP) to achieve better feature modeling for VTG. FDAP consists of two key com-

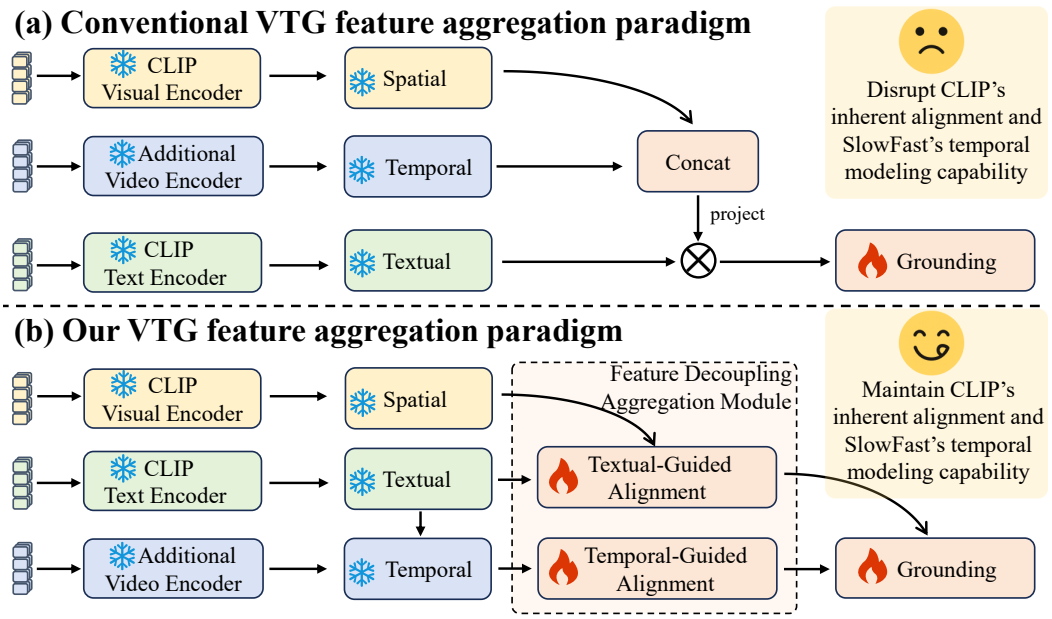

Figure 1: Comparison between the conventional feature modeling paradigm and our proposed paradigm for the VTG task. Existing methods adopt a "concat-then-project" aggregation approach, whereas our paradigm decouples this process for more effective feature modeling.

ponents: the Textual-Guided Feature Decoupling Module (TGFDM), which preserves the inherent alignment of CLIP features and maintains the temporal modeling capacity of SlowFast features; and the Dual-branch Feature Aggregation Module (DFAM), which dynamically adjusts the preference between textual and temporal features at the feature aggregation stage in a sample-adaptive manner.

Specifically, TGFDM computes a textual-guided attention map between CLIP's visual and textual features and a temporal-guided attention map from SlowFast features. These attention maps are independently applied to the CLIP and SlowFast visual streams, enriching representations with textual and temporal cues. DFAM then adaptively integrates the two guided features with sample-specific weights, enabling the model to dynamically adjust the preference for textual and temporal features during the feature aggregation stage.

To validate the effectiveness of our proposed FDAP, we integrate FDAP into four widely adopted VTG methods: M-DETR Lei et al. (2021), TR-DETR Sun et al. (2024), CG-DETR Moon et al. (2023a), and Flash-VTG Cao et al. (2025), and conduct extensive experiments across three datasets: QVHighlights Lei et al. (2021), Charades-STA Gao et al. (2017), and TACoS Regneri et al. (2013). The results demonstrate consistent performance improvements across all methods, *e.g.*, it improves R1@0.7 of M-DETR by about 3%. Notably, FDAP is highly lightweight, introducing only 0.2M additional parameters.

Our main contributions are summarized as follows:

- We propose the FDAP to tackle three critical issues in VTG feature modeling: the disruption of CLIP's cross-modal alignment, the degradation of SlowFast's temporal modeling capacity, and the neglect of query-dependent preferences between textual and temporal features.
- We design a TGFDM to preserve CLIP's cross-modal alignment while retaining SlowFast's temporal modeling capabilities. Additionally, we introduce a DFAM to dynamically adjust feature preference between textual and temporal features in a sample-adaptive manner.
- Extensive experiments on four VTG methods across three benchmarks demonstrate the effectiveness and generalizability of FDAP.

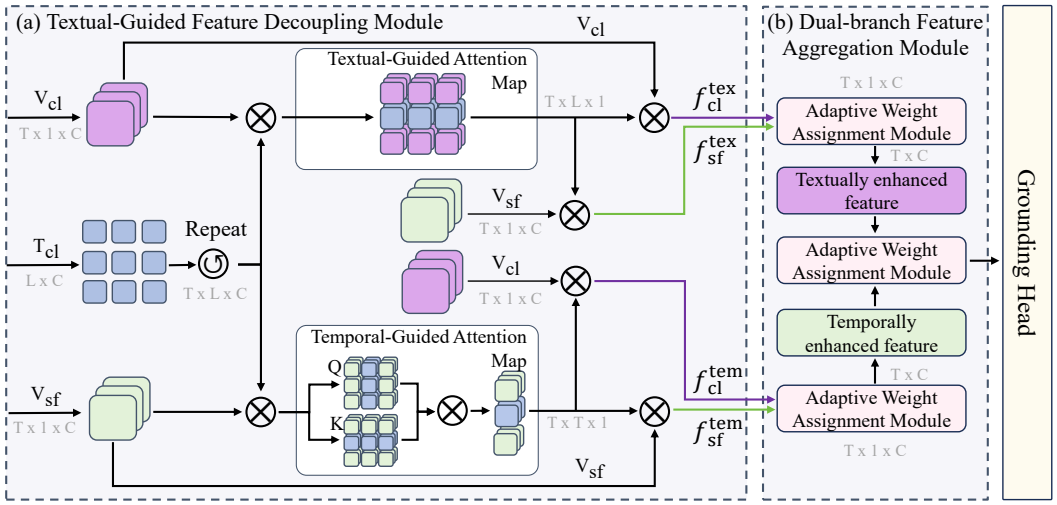

Figure 2: Overview of the Feature Decoupling and Aggregation Paradigm, which consists of two main components: (a) the Textual-Guided Feature Decoupling Module and (b) the Dual-Branch Feature Aggregation Module.

2 METHODOLOGY

2.1 OVERVIEW

In the Video Temporal Grounding (VTG) task, the objective is to localize the most relevant temporal segment within an untrimmed video V, given a natural language query Q. Specifically, the input video is processed by a frozen CLIP and SlowFast encoder to extract two parallel streams of visual features, which are then projected into a unified feature space of dimension C via a dedicated MLP layer, resulting in $\mathrm{V_{cl}}$ and $\mathrm{V_{sf}}$. Meanwhile, the textual features extracted from the CLIP text encoder are similarly projected and expanded as needed, yielding $\mathrm{T_{cl}}$. These multi-modal features are subsequently aggregated via simple concatenation and passed to a unified grounding head Φ, which performs two tasks: moment retrieval (MR) and highlight detection (HD). The MR head predicts the start and end timestamps of the target segment, while the HD head generates a continuous saliency score to identify frames that are most semantically aligned with the query.

However, under the conventional "concat-then-project" aggregation paradigm, the dual visual features $\mathrm{V_{cl}}$ and $\mathrm{V_{sf}}$ are directly concatenated and projected into a fixed dimension C, followed by cross-modal alignment with the textual features extracted by CLIP. This strategy inevitably disrupts the inherent cross-modal alignment originally established by CLIP, while also undermining the temporal modeling capabilities introduced by SlowFast. To mitigate the misalignment caused by naive feature concatenation, we propose a Feature Decoupling and Aggregation Paradigm (FDAP), which is composed of two key modules: the Textual-Guided Feature Decoupling Module (TGFDM) and the Dual-branch Feature Aggregation Module (DFAM).

Specifically, TGFDM preserves CLIP's original cross-modal alignment by decoupling the dual visual encoders and generating two types of enhanced visual representations: textually enhanced features ($f_{\mathrm{cl}}^{\mathrm{tex}}, f_{\mathrm{sf}}^{\mathrm{tex}}$) and temporally enhanced features ($f_{\mathrm{cl}}^{\mathrm{tem}}, f_{\mathrm{sf}}^{\mathrm{tem}}$), formulated as follows:

$$\left\{f_{\mathrm{cl}}^{\mathrm{tex}}, f_{\mathrm{cl}}^{\mathrm{tem}}, f_{\mathrm{sf}}^{\mathrm{tex}}, f_{\mathrm{sf}}^{\mathrm{tem}}\right\} = \mathrm{TGFDM}(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{sf}}) \tag{1}$$

Considering that different queries often exhibit different preferences for textual and temporal features in VTG, it is crucial to design an aggregation strategy that accounts for such query-dependent characteristics. However, conventional aggregation paradigms typically overlook such query-dependent preferences. To address this limitation, we introduce the Dual-branch Feature Aggregation Module (DFAM), which incorporates three independent Adaptive Weight Aggregation Modules

(AWAM) to compute sample-specific aggregation weights in a query-adaptive manner. This design enables the model to effectively fuse the four enhanced visual representations based on the textual and temporal preferences implied by each query. Formally, the aggregation is defined as:

$$\hat{\mathrm{F}} = \Phi\big(\mathrm{DFAM}(f_{\mathrm{cl}}^{\mathrm{tex}}, f_{\mathrm{sf}}^{\mathrm{tex}}, f_{\mathrm{cl}}^{\mathrm{tem}}, f_{\mathrm{sf}}^{\mathrm{tem}})\big) \tag{2}$$

The final representation is routed to the task-specific head Φ to produce $\hat{\mathrm{F}}$ for label prediction, supporting the downstream tasks of Moment Retrieval and Highlight Detection.

2.2 Textual-Guided Feature Decoupling Module

To address the three critical issues in feature modeling, we propose the Textual-Guided Feature Decoupling Module (TGFDM), which specifically targets the first two: preserving the inherent cross-modal alignment of CLIP features and maintaining the temporal modeling capacity of SlowFast features, as illustrated in Figure 2. To achieve this, TGFDM computes two types of enhanced visual representations using a textual-guided attention mechanism $\mathrm{Att}_{\mathrm{tex}}$ and a temporal-guided attention mechanism $\mathrm{Att}_{\mathrm{tem}}$, yielding:

$$\begin{aligned}\left\{f_{\mathrm{cl}}^{\mathrm{tex}}, f_{\mathrm{sf}}^{\mathrm{tex}}, f_{\mathrm{cl}}^{\mathrm{tem}}, f_{\mathrm{sf}}^{\mathrm{tem}}\right\} &= \mathrm{TGFDM}\left(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{sf}}\right) \\ &= \begin{cases} \mathrm{Att_{tex}}\left(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{cl}}\right) \\ \mathrm{Att_{tex}}\left(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{sf}}\right) \\ \mathrm{Att_{tem}}\left(\mathrm{T_{cl}}, \mathrm{V_{sf}}, \mathrm{V_{cl}}\right) \\ \mathrm{Att_{tem}}\left(\mathrm{T_{cl}}, \mathrm{V_{sf}}, \mathrm{V_{sf}}\right) \end{cases}\end{aligned} \tag{3}$$

We decompose the procedure as follows: First, to preserve the original cross-modal alignment established by CLIP's visual and textual encoders, we introduce a textual-guided attention mechanism, which computes a textual-guided attention map $\mathcal{M}_{\mathrm{cl}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{L}\times 1}$ based on the visual and textual features extracted from CLIP, where T denotes the total number of video frames, with L being the number of tokens in the query. This map captures the semantic relevance between video content and the query. The attention map $\mathcal{M}_{\mathrm{cl}}$ is then applied to both the CLIP and SlowFast visual features to obtain the corresponding textually enhanced visual representations. The computation is formally defined as:

$$\begin{aligned}\mathcal{M}_{\mathrm{cl}} &= \mathrm{Softmax}\left(\frac{\mathrm{T_{cl}} \cdot \mathrm{V_{cl}}^{T}}{\sqrt{\mathrm{C}}}\right) \\ f_{\mathrm{cl}}^{\mathrm{tex}} &= \mathrm{Att_{tex}}\left(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{cl}}\right) \\ &= \mathcal{M}_{\mathrm{cl}} \cdot \mathrm{V_{cl}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{L}\times\mathrm{C}} \\ f_{\mathrm{sf}}^{\mathrm{tex}} &= \mathrm{Att_{tex}}\left(\mathrm{T_{cl}}, \mathrm{V_{cl}}, \mathrm{V_{sf}}\right) \\ &= \mathcal{M}_{\mathrm{cl}} \cdot \mathrm{V_{sf}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{L}\times\mathrm{C}}\end{aligned} \tag{4}$$

Here, T represents the matrix transpose operation. By using $\mathcal{M}_{\mathrm{cl}}$ to attend over the CLIP features $\mathrm{V_{cl}}$, we obtain $f_{\mathrm{cl}}^{\mathrm{tex}}$; similarly, by replacing the value component with the SlowFast features $\mathrm{V_{sf}}$, we derive the textually enhanced SlowFast representation, denoted as $f_{\mathrm{sf}}^{\mathrm{tex}}$.

Second, to address CLIP's limited ability to capture inter-frame dependencies, we incorporate the SlowFast network into our framework and introduce a temporal-guided attention mechanism to model query-dependent temporal correlations. Analogous to the textual-guided attention mechanism, this approach differs by explicitly capturing frame-level dependencies through computing the temporal-guided attention map $\mathcal{M}_{\mathrm{sf}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}\times 1}$ along the temporal dimension. The computation is formally defined as follows:

$$\begin{aligned}\mathcal{A} &= \mathrm{T_{cl}} \cdot \mathrm{V_{sf}}^{T} \\ \mathcal{M}_{\mathrm{sf}} &= \mathrm{Softmax}\left(\frac{\mathcal{A} \cdot \mathcal{A}^{T}}{\sqrt{\mathrm{C}}}\right) \\ f_{\mathrm{cl}}^{\mathrm{tem}} &= \mathrm{Att_{tem}}\left(\mathrm{T_{cl}}, \mathrm{V_{sf}}, \mathrm{V_{cl}}\right) \\ &= \mathcal{M}_{\mathrm{sf}} \cdot \mathrm{V_{cl}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}\times\mathrm{C}} \\ f_{\mathrm{sf}}^{\mathrm{tem}} &= \mathrm{Att_{tem}}\left(\mathrm{T_{cl}}, \mathrm{V_{sf}}, \mathrm{V_{sf}}\right) \\ &= \mathcal{M}_{\mathrm{sf}} \cdot \mathrm{V_{sf}} \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}\times\mathrm{C}}\end{aligned} \tag{5}$$

Table 1: Comparison with state-of-the-art methods on the Charades-STA and TACoS datasets. "*" denotes results under our FDAP. In the upper group, the highest score is underlined, while in the lower group, the highest score is **bolded**.

Method	Published	Charades-STA			TACoS		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
M-DETR	NeurIPS'21	–	–	–	24.67	11.97	25.49
2D-TAN	AAAI'20	46.02	27.50	–	27.99	12.92	27.22
VSLNet	ACL'20	47.31	30.19	45.15	23.54	13.15	24.99
QD-DETR	CVPR'23	57.31	32.55	–	–	–	–
UniVTG	ICCV'23	58.01	35.65	50.10	34.77	17.35	33.60
LLMEPET	MM'24	–	36.49	50.25	–	22.78	36.55
R^2-Tuning	ECCV'24	59.78	37.02	50.86	38.72	25.12	35.92
CG-DETR	arXiv'24	56.26 ± 0.41	32.82 ± 0.51	48.35 ± 0.28	37.88 ± 0.68	22.02 ± 0.39	35.99 ± 0.94
CG-DETR*	Ours	$\mathbf{57.37 \pm 0.24}$	$\mathbf{34.78 \pm 0.50}$	$\mathbf{49.39 \pm 0.33}$	$\mathbf{38.91 \pm 0.40}$	$\mathbf{23.42 \pm 0.58}$	$\mathbf{36.27 \pm 0.39}$
Flash-VTG	WACV'25	57.86 ± 0.75	35.45 ± 0.30	49.63 ± 0.50	39.88 ± 0.54	25.73 ± 0.52	36.56 ± 0.55
Flash-VTG*	Ours	$\mathbf{59.82 \pm 0.46}$	$\mathbf{37.27 \pm 0.97}$	$\mathbf{51.11 \pm 0.17}$	$\mathbf{40.74 \pm 0.48}$	$\mathbf{26.40 \pm 0.43}$	$\mathbf{37.16 \pm 0.34}$

where $\mathcal{A}$ denotes the attention logits reflecting the semantic correlation between the CLIP's textual query and SlowFast's visual features.

2.3 Dual-branch Feature Aggregation Module

While prior studies leverage both textual and temporal features for video-query alignment, they often overlook query-dependent preferences between these modalities. Most existing methods apply uniform aggregation weights, ignoring sample-specific biases. To address this, we propose DFAM, which enables flexible and query-sensitive fusion via three independent AWAMs.

Adaptive Weight Aggregation Module. To facilitate query-dependent preference modeling and enable more effective aggregation of dual-stream representations, we propose the AWAM. In contrast to conventional strategies such as fixed-weight summation or naïve concatenation with shared global parameters, AWAM adaptively estimates weighting coefficients in a sample-specific manner, allowing the model to dynamically adjust its emphasis on textual or temporal features based on the input query, which is essential for the task of VTG. Concretely, we apply max pooling to the textual-guided attention maps $\mathcal{M}_{\text{cl}}$ and temporal-guided attention maps $\mathcal{M}_{\text{sf}}$ along the spatial (L) and temporal (T) dimensions, respectively, and concatenate the resulting descriptors. This concatenated representation is then fed into a parameter-independent multi-layer perception (MLP), followed by a Softplus activation, to produce the textual weight λ_1 and the temporal weight λ_2. Formally defined as:

$$\lambda_1, \lambda_2 = \text{MLP}\big(\text{Maxpool}(\mathcal{M}_{\text{cl}}) \,\|\, \text{Maxpool}(\mathcal{M}_{\text{sf}})\big) \tag{6}$$

where both inputs and outputs are tensors of shape $\mathbb{R}^{\text{T}\times 2}$, and $\|$ denotes the concatenation operation.

We subsequently perform textually and temporally enhanced feature aggregation using AWAM, denoted as σ. For textually enhanced feature, we compute a unified representation by applying the learned weights λ_1 and λ_2 to $f_{\text{cl}}^{\text{tex}}$ and $f_{\text{sf}}^{\text{tex}}$, respectively. The aggregation is formally defined as:

$$\sigma\left(f_{\text{cl}}^{\text{tex}}, f_{\text{sf}}^{\text{tex}}\right) = \lambda_1 \cdot f_{\text{cl}}^{\text{tex}} + \lambda_2 \cdot f_{\text{sf}}^{\text{tex}} \tag{7}$$

In parallel, temporally enhanced feature is produced by weighting $f_{\text{cl}}^{\text{tem}}$ and $f_{\text{sf}}^{\text{tem}}$ with the same coefficients. This adaptive aggregation strategy enables flexible and context-aware aggregation of dual-stream features at both semantic and temporal levels. Finally, we apply AWAM once more to aggregate two enhanced features into the final output $\hat{\text{F}}$. Here, the grounding head for downstream tasks is denoted by Φ. The complete aggregation process can be explicitly formulated as follows:

$$\begin{aligned}\hat{\text{F}} &= \Phi\Big(\text{DFAM}\big(f_{\text{cl}}^{\text{tex}}, f_{\text{sf}}^{\text{tex}}, f_{\text{cl}}^{\text{tem}}, f_{\text{sf}}^{\text{tem}}\big)\Big) \\ &= \Phi\Big(\sigma_{\text{glb}}\big(\sigma_{\text{tex}}(f_{\text{cl}}^{\text{tex}}, f_{\text{sf}}^{\text{tex}}), \sigma_{\text{tem}}(f_{\text{cl}}^{\text{tem}}, f_{\text{sf}}^{\text{tem}})\big)\Big)\end{aligned} \tag{8}$$

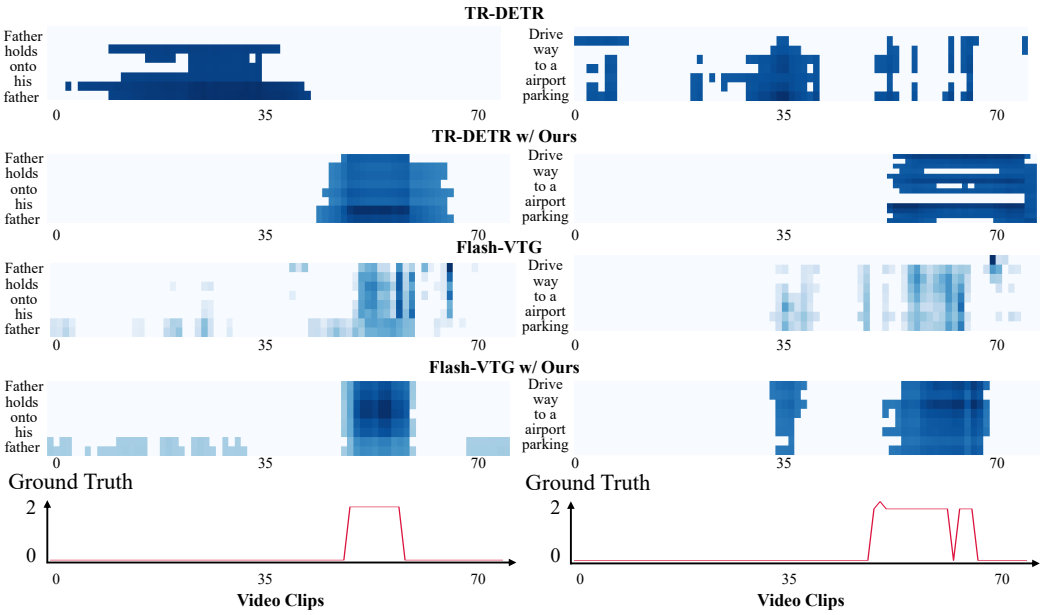

Figure 3: Visualization of the query-aware representation capacity of existing VTG methods under the conventional "concat-then-project" paradigm and our proposed FDAP.

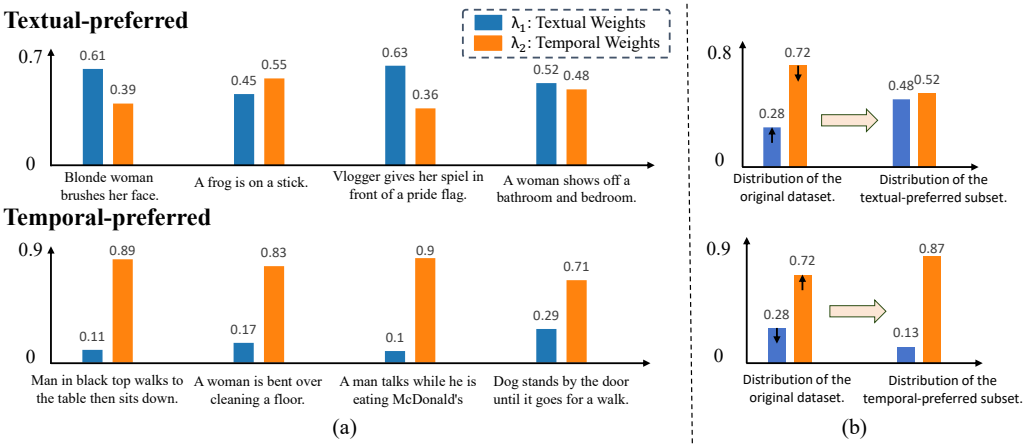

Figure 4: Visualization of query-dependent textual and temporal preferences in VTG, as reflected by the final AWAM weights across different queries. (a) Textual and temporal weights of representative samples. (b) Dataset-level distribution of textual and temporal weights.

Training and Inference Setting. Our FDAP serves as a head-agnostic and general feature modeling paradigm. As an end-to-end method, it can be integrated into various pipelines while preserving original experimental protocols and hyperparameter settings across different architectures.

Discussion. To qualitatively evaluate the model's query perception capability, Figure 3 visualizes the query-guided attention maps derived from the output representation $\hat{F}$, produced by TR-DETR and Flash-VTG under our FDAP, on the QVHighlights dataset. The attention map is presented alongside the corresponding ground-truth video segments that are most relevant to the given query. In the visualization, darker colors indicate higher relevance scores, while lighter colors represent lower relevance. As illustrated, our method localizes query-relevant segments more precisely and produces higher relevance scores compared to TR-DETR and Flash-VTG. These improvements can

Table 2: Comparison on the QVHighlights validation set. "†" denotes methods using audio; "*" indicates results under our FDAP. In the upper group, the highest score is underlined, while in the lower group, the highest score is **bolded**.

Method	Published	Moment Retrieval			Highlight Detection	
		R1@0.5	R1@0.7	mAP	mAP	Hit@1
UniVTG	ICCV'23	59.74	–	38.83	36.13	61.81
UMT†	CVPR'22	60.26	44.26	38.59	39.85	64.19
EaTR	ICCV'23	61.36	45.79	41.74	37.15	58.61
QD-DETR	CVPR'23	62.68	46.66	41.22	39.83	63.05
TaskWeave	CVPR'24	64.26	50.06	45.38	39.28	63.68
UVCOM	CVPR'24	65.10	51.81	45.79	–	–
LLMEPET	MM'24	66.58	51.10	46.24	40.52	65.03
R^2-Tuning	ECCV'24	68.71	52.06	47.59	40.59	64.32
M-DETR	NeurIPS'21	52.97 ± 1.13	32.88 ± 2.29	30.87 ± 1.35	35.84 ± 0.38	55.29 ± 0.68
M-DETR*	Ours	**55.33 ± 1.43**	**35.77 ± 1.37**	**32.47 ± 1.22**	**36.30 ± 0.56**	**56.59 ± 1.93**
TR-DETR	AAAI'24	65.78 ± 0.78	49.67 ± 0.86	43.71 ± 0.93	39.43 ± 0.25	64.43 ± 0.39
TR-DETR*	Ours	**66.00 ± 0.60**	**50.66 ± 1.13**	**44.38 ± 0.58**	**40.42 ± 0.20**	**64.85 ± 0.69**
CG-DETR	arXiv'24	64.88 ± 0.38	49.97 ± 0.39	43.67 ± 0.51	39.95 ± 0.34	64.42 ± 0.51
CG-DETR*	Ours	**66.28 ± 0.52**	**51.33 ± 0.89**	**45.15 ± 0.37**	**40.43 ± 0.31**	**65.20 ± 1.05**
Flash-VTG	WACV'25	67.78 ± 0.78	52.36 ± 0.88	48.92 ± 0.53	41.08 ± 0.29	66.01 ± 0.74
Flash-VTG*	Ours	**68.66 ± 0.60**	**53.73 ± 0.42**	**49.46 ± 0.46**	**41.35 ± 0.13**	**67.37 ± 0.98**

be attributed to the decoupled design of FDAP, which enables the model to effectively leverage both textually enhanced and temporally enhanced representations, thereby facilitating more precise grounding of query-dependent video segments.

Beyond analyzing how the aggregated features respond to query semantics, understanding the inherent preference of queries toward different features is equally essential yet remains underexplored. Figure 4 visualizes the textual and temporal preferences captured by the final AWAM module of Flash-VTG under our FDAP setting on the QVHighlights dataset, where (a) presents the textual and temporal weights of representative samples and (b) depicts their distributions across the original dataset as well as the subsets derived from our preference-based partitioning. The classification is guided by linguistic cues such as the presence of temporal connectives (while, then, until, e.g.), the number of event verbs, and the dominance of stative verbs or relational constructs, resulting in 3,415 textual-preferred queries and 3,803 temporal-preferred queries. For instance, the query "A man talks while he is eating McDonald's" contains the temporal connective (while) and two dynamic actions (talk, eat), thus categorized as temporal-preferred, whereas "Kids are sitting on the floor" lacks temporal connectives and involves the stative verb (sit), aligning with the textual-preferred category.

Notably, we compute the weighting coefficients λ_1 and λ_2 for each frame and average them across all frames to obtain the final sample-adaptive aggregation weights. As shown, queries dominated by nouns tend to rely more on textual features, whereas those involving states or actions are more influenced by temporal features when identifying the relevant video segments.

3 Experiments

3.1 Dataset and Evaluation Metrics

In this paper, we conduct comprehensive experiments on three widely used benchmarks for the Video Temporal Grounding (VTG) task: QVHighlights Lei et al. (2021), Charades-STA Gao et al. (2017), and TACoS Regneri et al. (2013). Among them, QVHighlights is the most extensively adopted, as it provides annotations for both MR and HD. The dataset contains over 10,000 daily vlogs and news videos paired with natural language queries, where each video may include multiple moment-level retrieval clips. This setting better reflects real-world scenarios and retrieval challenges. In addition, we evaluate our model on Charades-STA and TACoS for MR, which focus on indoor daily activities and complement grounding performance in structured environments.

Table 3: Ablation study on different components conducted within the Flash-VTG framework on the Charades-STA dataset. The highest score in each row is **bolded**.

Component	+ DFAM	✓		✓
	+ TGFDM		✓	✓
Metrics	R1@0.3	68.25	69.14	**71.32 ± 0.40**
	R1@0.5	54.89	58.33	**59.82 ± 0.46**
	R1@0.7	32.77	35.53	**37.27 ± 0.97**
	mAP	37.27	40.03	**41.35 ± 0.26**
	mAP@0.5	65.56	67.63	**69.39 ± 0.36**
	mAP@0.75	36.22	39.85	**41.16 ± 0.50**

We follow the standard evaluation protocols adopted in previous studies. For HD, we employ mAP and HIT@1 as evaluation metrics, where a prediction is considered correct if it receives a saliency score of Very Good. For MR, we report R1 at IoU thresholds of 0.5 and 0.7, mAP at the same thresholds, and the average mAP computed over a range of IoU thresholds from 0.5 to 0.95 with an interval of 0.05 on the QVHighlights dataset. Additionally, we assess MR performance on Charades-STA and TACoS using R1 at IoU thresholds of 0.3, 0.5, and 0.7, as well as the mIoU.

3.2 Implementation Details

In all experiments, we utilize the same visual and textual features extracted by CLIP Radford et al. (2021) and SlowFast Feichtenhofer et al. (2018) across the three datasets to ensure fair comparisons. To mitigate variance from random seeds, we follow M-DETR's protocol and report mean ± std over five runs. We evaluate our method on four representative benchmarks: M-DETR Lei et al. (2021), TR-DETR Sun et al. (2024), CG-DETR Moon et al. (2023a), and Flash-VTG Cao et al. (2025). Since the official implementations of M-DETR and TR-DETR are not publicly available on Charades-STA and TACoS, these two baselines are only evaluated on the QVHighlights dataset. All experiments are conducted using the same five random seeds on a single RTX 4090 GPU. Other experimental settings and hyperparameters strictly follow the original configurations. For parameter and FPS comparisons in Table 5, the batch size is fixed to 32.

3.3 Comparison Results

We first evaluate the effectiveness of the proposed FDAP on four benchmark datasets, beginning with Charades-STA and TACoS, which are widely adopted in VTG and primarily represent in-domain scenarios. The corresponding results are reported in Table 1. As shown, integrating FDAP consistently improves performance across all evaluation metrics, including R1@0.5, R1@0.7, and mIoU, compared to the baselines. Notably, in both Table 1 and Table 2, the upper group of results corresponds to those reported in the original papers, while the lower group presents our re-implemented results, averaged over five runs with fixed random seeds following the M-DETR evaluation protocol (mean ± standard deviation). Specifically, incorporating FDAP into Flash-VTG leads to a 1.96% improvement in R1@0.5 and a 1.48% gain in mIoU. Beyond performance improvements, FDAP also contributes to improved result stability. In the lower group of Table 1, 83.3% of the metrics exhibit reduced standard deviation, demonstrating the robustness of the proposed paradigm. These results highlight the effectiveness of FDAP in enhancing accuracy and stability in moment retrieval.

To validate our method for highlight detection, we conduct additional experiments on the QVHighlights validation set, which supports both MR and HD tasks. The results for both tasks are summarized in Table 2. Our FDAP approach consistently improves performance across all baseline methods on MR metrics, achieving gains of approximately 0.5%–1% over TR-DETR and Flash-VTG, around 2% over CG-DETR, and nearly 3% over M-DETR. In addition, we assess the performance of FDAP on the HD task using HL-mAP and HL-HIT@1. As shown in Table 2, Flash-VTG with FDAP yields a 1.36% improvement in HL-HIT@1; when combined with CG-DETR, it achieves a 0.78% gain in HL-HIT@1; and with TR-DETR, it brings a 0.99% improvement in HL-mAP. These results further demonstrate the robustness and effectiveness of our approach.

Table 4: Ablation study of different feature aggregation methods conducted using Flash-VTG on the Charades-STA dataset. The highest score in each column is **bolded**.

Setting	R1@0.5	R1@0.7	mIoU
Concat	57.53	34.33	49.51
Add	58.33	35.53	49.67
Weighted Sum	58.04	36.4	50.06
AWAM (Ours)	**59.82 ± 0.46**	**37.27 ± 0.97**	**51.11 ± 0.17**

Table 5: Ablation study on FPS and Parameters of different methods with our FDAP on the QVHighlights dataset. "*" denotes models equipped with our FDAP.

Metric	M-DETR	TR-DETR	CG-DETR	Flash-VTG
FPS	498.91	281.97	39.41	175.59
Params (M)	4.6	7.9	12.0	10.6
Metric	**M-DETR***	**TR-DETR***	**CG-DETR***	**Flash-VTG***
FPS	469.78	265.29	36.43	159.67
Params (M)	4.8	8.1	12.2	10.8

3.4 Ablation Study

As shown in Table 3, we analyze the individual contributions of the TGFDM and the DFAM. Notably, the TGFDM yields a performance gain over the baseline, with both mAP and R1 score for MR improving by nearly 3%. In addition, the DFAM further contributes to performance improvement, leading to an additional gain of nearly 2%.

Within our AWAM, we observe that different feature aggregation strategies significantly affect model performance. In addition to the qualitative analysis presented in Figure 4, we further provide a quantitative comparison of different aggregation methods in Table 4. To this end, we conduct a comprehensive evaluation on the Charades-STA dataset, comparing conventional strategies including concatenation, element-wise addition, and weighted sum against our proposed AWAM. The evaluation is performed using R1@0.5, R1@0.7, and mIoU metrics. As shown in Table 4, AWAM outperforms the commonly adopted "concat-then-project" paradigm, yielding improvements of 2.29% and 2.94% in R1@0.5 and R1@0.7, respectively, along with a 1.6% gain in mIoU. These results demonstrate the effectiveness of AWAM for feature aggregation in the decoupling framework.

Moreover, our FDAP introduces negligible computational overhead, introducing merely 0.2M additional parameters and incurring a 5%–9% increase in inference time (FPS), as shown in Table 5. For example, when integrated into Flash-VTG, the additional parameters account for only 1.85% of the total model size. Specifically, FDAP increases FPS latency by 5.83%, 5.91%, 7.56%, and 9.06% on M-DETR, TR-DETR, CG-DETR, and Flash-VTG, respectively. This lightweight design enables efficient deployment in resource-constrained scenarios with negligible computational cost.

4 Conclusion

In this paper, we propose FDAP, a plug-and-play Feature Decoupling and Aggregation Paradigm for Video Temporal Grounding. Unlike the conventional "concat-then-project" paradigm, FDAP decouples the CLIP and SlowFast features and employs a Textual-Guided Feature Decoupling Module to maintain SlowFast's temporal modeling capability, while preserving the original alignment between CLIP's visual and textual representations. Furthermore, we introduce DFAM to adjust the preference between textual and temporal features in a sample-adaptive manner. Extensive experiments on three VTG benchmarks demonstrate the superior performance of FDAP on both the Moment Retrieval and Highlight Detection tasks. We hope that our FDAP can inspire further exploration into feature aggregation strategies for VTG tasks within the community.

Ethics and Reproducibility Statement

This work poses no ethical concerns, as it does not involve human subjects, sensitive data, or potentially harmful applications. The datasets used are described in Section 3.1, and detailed parameter settings are provided in Section 3.2. The complete source code will be released upon acceptance of the paper.

References

Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. FlashVTG: Feature Layering and Adaptive Score Handling Network for Video Temporal Grounding . In *WACV*, pp. 9226–9236, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/WACV61041.2025.00894. URL https://doi.ieeecomputersociety.org/10.1109/WACV61041.2025.00894.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *ICCV*, pp. 6201–6210, 2018. URL https://api.semanticscholar.org/CorpusID:54463801.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563.

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *CVPR*, pp. 6565–6574, 2023. doi: 10.1109/CVPR52729.2023.00635.

Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, pp. 13800–13810, 2023. doi: 10.1109/ICCV51070.2023.01273.

Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *MM*, MM '24, pp. 7249–7258, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681115. URL https://doi.org/10.1145/3664647.3681115.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pp. 105–124, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19832-8. doi: 10.1007/978-3-031-19833-5_7. URL https://doi.org/10.1007/978-3-031-19833-5_7.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *NeurIPS*, NeurIPS, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, pp. 2782–2792, 2023. doi: 10.1109/ICCV51070.2023.00262.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pp. 3032–3041, 2022. doi: 10.1109/CVPR52688.2022.00305.

Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *ECCV*, pp. 421–438, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72939-3. doi: 10.1007/978-3-031-72940-9_24. URL https://doi.org/10.1007/978-3-031-72940-9_24.

WonJun Moon, Sangeek Hyun, Subeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *ArXiv*, abs/2311.08835, 2023a. URL https://api.semanticscholar.org/CorpusID:265213284.

WonJun Moon, Sangeek Hyun, Sang shin Paldal-gu Suwon-city Park, Dongchan Park, and Jae-Pil Heo. Query - dependent video representation for moment retrieval and highlight detection. *CVPR*, pp. 23023–23033, 2023b. URL https://api.semanticscholar.org/CorpusID:257757326.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 03 2013. ISSN 2307-387X. doi: 10.1162/tacl_a_00207. URL https://doi.org/10.1162/tacl_a_00207.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL https://api.semanticscholar.org/CorpusID:14124313.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: task-reciprocal transformer for joint moment retrieval and highlight detection. In *AAAI*. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28304. URL https://doi.org/10.1609/aaai.v38i5.28304.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, pp. 18709–18719, 2024. doi: 10.1109/CVPR52733.2024.01770.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *ICCV*, pp. 13577–13587, 2023. doi: 10.1109/ICCV51070.2023.01253.

Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *CVPR*, pp. 18308–18318, 2024. doi: 10.1109/CVPR52733.2024.01733.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6543–6554, Online, July 2020a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.585.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pp. 12870–12877, 2020b.

Henghao Zhao, Kevin Qinghong Lin, Rui Yan, and Zechao Li. Diffusionvmr: Diffusion model for joint video moment retrieval and highlight detection. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024. doi: 10.1109/TNNLS.2024.3516033.

THE USE OF LARGE LANGUAGE MODELS

Large language models (LLMs) are employed solely for language refinement and are not utilized for information retrieval, knowledge discovery, or research ideation.

CONTENTS OF APPENDIX

This supplementary material provides additional discussions and experiments to further support the effectiveness of our proposed method. The appendix is organized into three parts:

A) Related Work. We review prior research on Video Temporal Grounding (VTG), from early benchmarks on moment retrieval and highlight detection Regneri et al. (2013); Gao et al. (2017) to unified benchmarks such as QVHighlights Lei et al. (2021). Advances in cross-modal feature modeling range from "CLIP-only" approaches Liu et al. (2024); Huang et al. (2023); Radford et al. (2021) to "CLIP+" paradigms that incorporate temporal encoders like SlowFast Feichtenhofer et al. (2018), VGG Simonyan & Zisserman (2014), or C3D Carreira & Zisserman (2017), with representative works including TR-DETR Sun et al. (2024), CG-DETR Moon et al. (2023a), and Flash-VTG Cao et al. (2025). A key underexplored challenge is query-dependent feature preference, where different queries emphasize textual or temporal cues; our FDAP addresses this by decoupling features and dynamically adapting aggregation to query semantics.

B) Extended Visualization Results. We present additional qualitative comparisons on QVHighlights and Charades-STA, evaluating TR-DETR, CG-DETR, and Flash-VTG alongside their FDAP-enhanced counterparts. As illustrated in Figures 5 and 6, the examples include both single and multiple relevant segments. For fair comparison, all predicted score sequences are normalized by their respective maximum values. Across both datasets, FDAP consistently produces predictions that better align with the ground-truth score distributions and exhibit stronger saliency responses, demonstrating its effectiveness and generalizability in capturing query-relevant moments under complex temporal conditions.

C) Additional Quantitative Results. We provide further quantitative comparisons for M-DETR, TR-DETR, CG-DETR, and Flash-VTG. Specifically, we report detailed metrics on the QVHighlights validation and test sets, as well as on the Charades-STA and TACoS datasets, averaged over five runs with fixed random seeds. This offers a comprehensive comparison between the original models and their FDAP-enhanced versions.

A RELATED WORK

Video Temporal Grounding. Video Temporal Grounding (VTG) aims to localize semantically relevant temporal segments within untrimmed videos conditioned on natural language queries. The task comprises several subtasks, such as moment retrieval (MR) and highlight detection (HD), all of which require precise cross-modal alignment between visual content and linguistic semantics over time. While early benchmarks Regneri et al. (2013); Gao et al. (2017) addressed MR and HD as separate problems, recent benchmarks like QVHighlights Lei et al. (2021) unify these subtasks to enable comprehensive evaluation, introducing Moment-DETR as a strong baseline for joint modeling that has laid the foundation for subsequent frameworks in VTG.

Cross-Modal Feature Modeling. In the task of VTG, feature modeling remains a central challenge, aiming to effectively capture and integrate features across diverse modalities. "CLIP-Only" methods, such as R2-Tuning Liu et al. (2024) and VoP Huang et al. (2023), leverage a single visual-text encoder—CLIP Radford et al. (2021)—to exploit the representational capacity of a unified backbone. These approaches focus on aligning visual and textual modalities within a shared embedding space. However, the modeling capacity of a single encoder remains limited. To address these limitations, recent research has shifted toward multi-modal feature aggregation, in which CLIP features are combined with temporal encoders such as SlowFast Feichtenhofer et al. (2018), VGG Simonyan & Zisserman (2014), or C3D Carreira & Zisserman (2017). As a result, several "CLIP+" approaches have emerged. For instance, TR-DETR Sun et al. (2024) introduces a local-global multi-modal alignment mechanism to project features from different modalities into a unified embedding space. CG-DETR Moon et al. (2023a) employs dummy tokens and adaptive cross-attention to suppress irrelevant video segments conditioned on the query. More recently, Flash-VTG Cao et al. (2025) proposes a Temporal Feature Layering module to model temporal dynamics, along with an Adaptive Score Refinement strategy that incorporates contextual cues to enhance frame-level relevance. However, recent works that adopt the "concat-then-project" aggregation paradigm disrupt the origi-

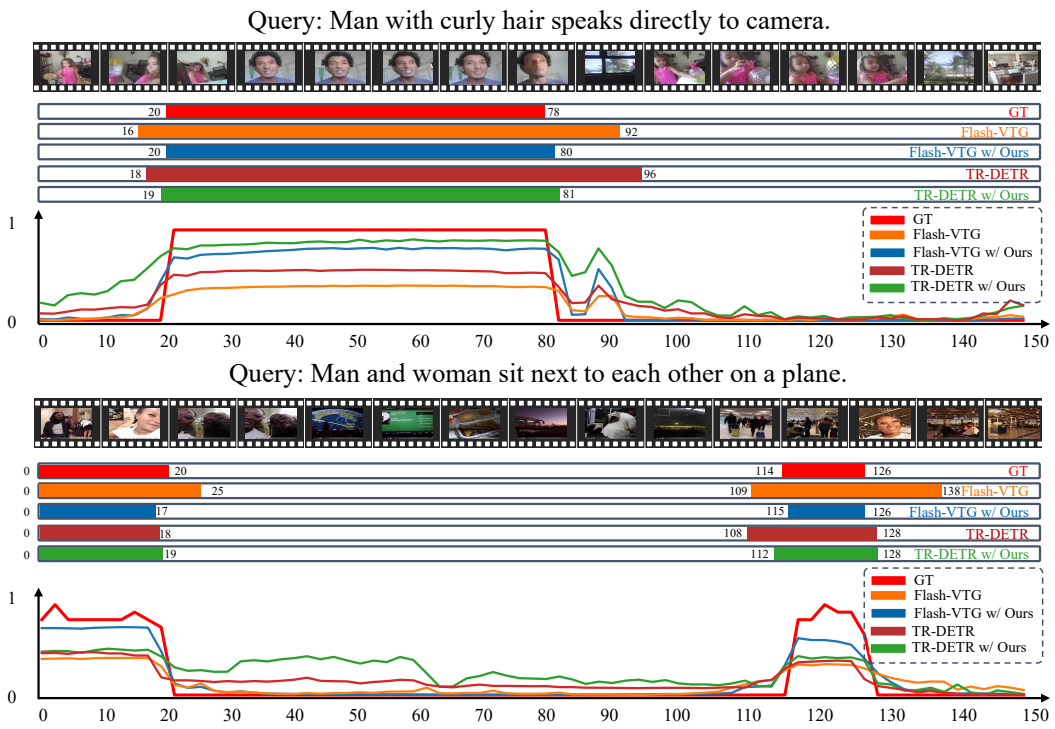

Figure 5: Qualitative results of TR-DETR and Flash-VTG, as well as their enhancements with our proposed FDAP, on the QVHighlights validation set.

nal alignment between CLIP's visual and textual representations and degrade the temporal modeling capability of SlowFast. FDAP effectively addresses the inherent limitations of the conventional aggregation paradigm through feature decoupling.

Query-Dependent Feature Preferences. Another underexplored challenge in VTG is the query-dependent preference for different feature modalities. Depending on the semantics of the query, some samples may benefit more from temporal cues (e.g., actions), while others rely heavily on appearance or contextual information. In conventional aggregation paradigms, textual and temporal features are processed jointly using shared parameters, which implicitly ignores such query-specific preferences. To address this limitation, our FDAP introduces the DFAM, which dynamically learns preference weights between textual and temporal features in a sample-adaptive manner. This design enhances the model's domain adaptability while preserving the original alignment between visual and textual representations.

B Extended Visualization Results

We present qualitative analysis results on both QVHighlights and Charades-STA to compare baseline models (TR-DETR, CG-DETR, and Flash-VTG) with their FDAP-enhanced counterparts. As shown in Figures 5 and 6, the examples cover cases with single and multiple relevant segments. For fair comparison, all predicted score sequences are normalized by their respective maximum values. Across both datasets, FDAP-augmented models (green and blue) yield predictions that align more closely with the ground-truth score distributions and exhibit higher saliency responses, compared to the original models (brown and orange). These consistent improvements highlight the effectiveness and generalizability of FDAP in accurately capturing query-relevant moments and enhancing robustness under complex temporal conditions.

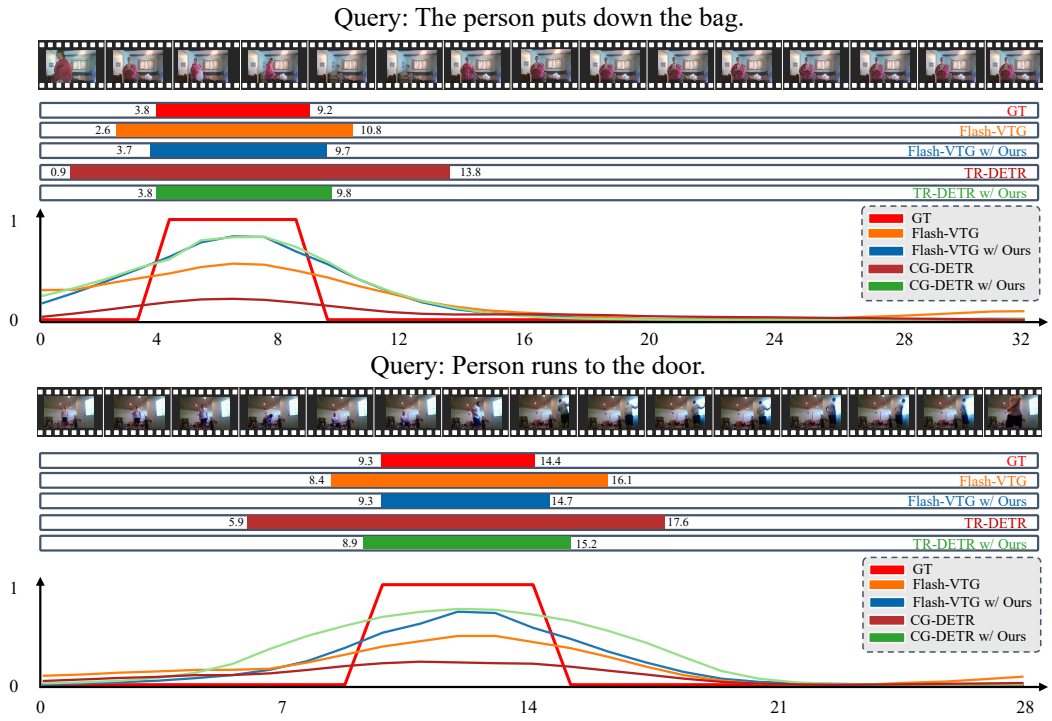

Figure 6: Qualitative results of CG-DETR and Flash-VTG, as well as their enhancements with our proposed FDAP, on the Charades-STA test set.

Table 6: Comparison results of M-DETR on the QVHighlights validation set. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
M-DETR-S1	53.10	34.13	31.38	54.51	30.14
M-DETR-S2	54.71	34.58	31.67	55.49	29.82
M-DETR-S3	52.06	32.13	29.81	53.61	27.77
M-DETR-S4	52.71	35.55	31.37	54.40	30.53
M-DETR-S5	52.13	33.42	30.09	53.51	27.84
M-DETR-S1*	54.84	36.45	32.19	55.28	30.48
M-DETR-S2*	54.39	35.87	32.05	55.28	31.71
M-DETR-S3*	57.68	37.61	34.55	58.31	33.42
M-DETR-S4*	55.61	34.77	32.24	56.23	31.11
M-DETR-S5*	53.10	35.55	31.50	54.40	30.17

C Additional Quantitative Results

Table 6 provides a detailed comparison of M-DETR, TR-DETR, CG-DETR, Flash-VTG with and without our proposed FDAP on the QVHighlights test and validation sets. All experiments are conducted under five independent runs with fixed random seeds (S1 to S5, corresponding to seed values 604, 2001, 2025, 4721, and 7717, respectively), ensuring statistical reliability of the reported results.

We observe that, even under the same model architecture, variations in random seed can lead to significant performance fluctuations. For example, in Table 6, the original M-DETR exhibits a large variance of **5.26%** in R1@0.7 on the test set, ranging from 28.92% (S3) to 34.18% (S2). With the

Table 7: Comparison results of TR-DETR on the QVHighlights validation set. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
TR-DETR-S1	64.65	48.65	42.25	63.87	43.28
TR-DETR-S2	66.19	50.13	43.78	65.27	43.31
TR-DETR-S3	65.61	50.77	44.70	65.30	45.80
TR-DETR-S4	66.77	49.81	44.25	65.97	44.98
TR-DETR-S5	65.68	48.97	43.59	64.91	43.87
TR-DETR-S1*	64.97	49.03	43.52	65.15	44.27
TR-DETR-S2*	66.13	51.81	44.74	64.71	45.47
TR-DETR-S3*	66.52	50.65	45.04	65.82	45.67
TR-DETR-S4*	66.06	50.19	44.35	66.30	45.64
TR-DETR-S5*	66.32	51.61	44.24	65.45	45.47

Table 8: Performance comparison of CG-DETR on the QVHighlights validation set. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
CG-DETR-S1	65.23	49.94	43.30	64.82	44.19
CG-DETR-S2	64.26	49.35	43.22	63.89	43.85
CG-DETR-S3	64.84	50.00	44.37	64.96	45.55
CG-DETR-S4	64.90	50.39	44.05	64.41	45.34
CG-DETR-S5	65.16	50.19	43.39	64.52	44.09
CG-DETR-S1*	67.03	52.84	45.80	66.35	46.79
CG-DETR-S2*	66.26	50.77	45.10	65.89	45.86
CG-DETR-S3*	66.26	51.35	45.05	66.07	46.30
CG-DETR-S4*	65.55	51.03	44.96	65.21	45.99
CG-DETR-S5*	66.32	50.65	44.85	65.98	45.47

introduction of our FDAP, this variance is substantially reduced to **2.86%**, with R1@0.7 ranging from 33.98% (S4*) to 36.84% (S3*). To account for the remaining variability and rigorously assess the effectiveness of our approach, we adopt an evaluation protocol that reports the mean and standard deviation over five runs with different fixed seeds.

Table 6 reports retrieval performance improvements achieved by integrating FDAP on the validation set. For R1@0.5, the baseline model achieves scores ranging from 52.06% (S3) to 54.71% (S2), while FDAP-enhanced models improve this to 53.10% (S5*) and 57.68% (S3*), corresponding to gains of 1.04% to 2.97%. For R1@0.7, the baseline scores span 32.13% (S3) to 35.55% (S4), and FDAP raises these to 34.77% (S4*) and 37.61% (S3*), yielding improvements of 2.64% to 2.06%. The mAP metric exhibits similar gains, increasing from 29.81% (S3) to 34.55% (S3*) after applying FDAP.

Table 7 summarizes the impact of FDAP on TR-DETR's performance across multiple metrics on the QVHighlights validation set. Compared to the baseline models, FDAP consistently improves results across all evaluation metrics. For instance, R1@0.5 increases from 64.65% (S1) to 66.52% (S3*), and R1@0.7 improves from 48.65% (S1) to 51.81% (S2*). Mean average precision (mAP) also shows notable gains, rising from a range of 42.25%–44.70% to 43.52%–45.04%. Similar improvements are observed in mAP@0.5 and mAP@0.75, demonstrating enhanced localization accuracy. These results confirm that FDAP consistently enhances TR-DETR's retrieval and localization capabilities across different random initializations, highlighting its robustness and generalizability.

Table 9: Performance comparison of Flash-VTG on the QVHighlights validation set. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
Flash-VTG-S1	68.52	53.55	49.66	68.67	51.48
Flash-VTG-S2	67.23	51.23	48.83	68.73	51.37
Flash-VTG-S3	68.39	52.13	48.69	68.62	50.03
Flash-VTG-S4	66.71	52.84	48.57	67.16	50.67
Flash-VTG-S5	68.06	52.06	48.84	68.01	50.19
Flash-VTG-S1*	68.06	53.42	49.96	68.89	52.33
Flash-VTG-S2*	68.06	53.23	49.55	68.48	51.61
Flash-VTG-S3*	69.35	53.81	49.24	69.17	51.57
Flash-VTG-S4*	68.65	53.87	49.75	69.33	51.95
Flash-VTG-S5*	69.16	54.32	48.78	68.80	51.18

Table 10: Comparison results of CG-DETR on the Charades-STA and TACoS datasets. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	**Charades-STA**					**TACoS**				
	R1@0.5	R1@0.7	mAP	mAP@0.5	mIoU	R1@0.5	R1@0.7	mAP	mAP@0.5	mIoU
CG-DETR-S1	56.59	33.04	34.27	65.73	48.55	38.39	21.54	21.06	42.40	36.23
CG-DETR-S2	56.83	32.47	33.77	65.41	48.66	38.24	22.04	21.31	42.34	36.40
CG-DETR-S3	55.73	32.02	33.20	63.65	48.23	36.89	22.59	20.95	41.24	34.13
CG-DETR-S4	55.81	33.23	33.98	65.40	47.89	37.47	21.94	21.33	41.87	36.52
CG-DETR-S5	56.32	33.33	35.57	67.53	48.40	38.42	21.97	22.23	44.80	36.75
CG-DETR-S1*	57.10	35.11	35.77	66.70	49.35	38.84	23.74	22.23	43.16	36.06
CG-DETR-S2*	57.39	34.76	35.39	66.24	49.46	38.32	22.57	21.41	42.67	35.72
CG-DETR-S3*	57.74	34.68	35.41	66.41	49.92	38.94	23.07	22.07	43.14	36.61
CG-DETR-S4*	57.12	33.98	35.11	66.66	49.06	39.42	23.79	23.05	44.23	36.67
CG-DETR-S5*	57.50	35.38	35.88	67.61	49.19	39.02	23.94	22.89	44.08	36.27

Table 8 reports the performance of CG-DETR with and without FDAP integration on the QVHighlights validation set. The baseline R1@0.5 scores range from 64.26% (S2) to 65.23% (S1), which are further improved to 65.55%–67.03% with FDAP. For R1@0.7, FDAP consistently boosts performance from 49.35%–50.39% to 50.65%–52.84%. Similarly, mAP increases from a baseline range of 43.22%–44.37% to 44.85%–45.80% after applying FDAP. Improvements are also observed in mAP@0.5 (from 63.89%–64.96% to 65.21%–66.35%) and mAP@0.75 (from 43.85%–45.55% to 45.47%–46.79%). These consistent gains across all metrics and seeds demonstrate the effectiveness and robustness of FDAP in enhancing CG-DETR's temporal grounding performance.

Table 9 reports the evaluation of Flash-VTG with and without FDAP across five fixed random seeds on the QVHighlights validation set. The integration of FDAP consistently enhances key performance metrics. Specifically, R1@0.5 improves from a baseline range of 66.71%–68.52% to 68.06%–69.35%, while R1@0.7 increases from 51.23%–53.55% to 53.23%–54.32%. Similarly, mAP is improved from 48.57%–49.66% to 48.78%–49.96%. FDAP also brings consistent gains in mAP@0.5 (from 67.16%–68.73% to 68.48%–69.33%) and mAP@0.75 (from 50.03%–51.48% to 51.18%–52.33%). These consistent improvements across all seeds confirm the robustness and generalizability of FDAP when applied to Flash-VTG under varied initialization conditions.

Table 10 presents the performance comparison of CG-DETR with and without FDAP on the Charades-STA and TACoS datasets across five fixed random seeds. The incorporation of FDAP consistently improves performance across all metrics and datasets. For Charades-STA, R1@0.5 increases from a baseline range of 55.73%–56.83% to 57.10%–57.74%, and R1@0.7 improves

Table 11: Comparison results of Flash-VTG on the Charades-STA and TACoS datasets. "*" denotes results obtained with our proposed FDAP. S1 through S5 correspond to experiments conducted with fixed random seeds.

Method	Charades-STA					TACoS				
	R1@0.5	R1@0.7	mAP	mAP@0.5	mIoU	R1@0.5	R1@0.7	mAP	mAP@0.5	mIoU
Flash-VTG-S1	58.50	35.91	39.87	68.52	50.22	39.99	26.24	27.10	48.52	36.66
Flash-VTG-S2	58.15	35.40	39.22	67.45	49.56	39.04	25.22	26.34	47.78	35.65
Flash-VTG-S3	57.02	35.32	39.20	66.75	49.04	40.24	25.57	27.73	49.04	37.06
Flash-VTG-S4	57.60	34.22	39.10	67.96	49.29	39.72	25.29	27.73	48.65	36.52
Flash-VTG-S5	58.60	35.32	39.47	68.49	50.04	40.41	26.32	27.58	49.88	36.89
Flash-VTG-S1*	60.56	36.42	41.32	69.94	51.35	40.11	26.82	28.32	49.39	37.57
Flash-VTG-S2*	59.54	36.77	41.55	69.30	51.05	40.34	25.77	27.40	49.59	36.79
Flash-VTG-S3*	59.97	36.51	40.92	69.50	50.88	41.09	26.29	27.87	49.97	37.16
Flash-VTG-S4*	59.54	38.39	41.51	69.25	51.09	41.04	26.79	28.38	50.20	37.42
Flash-VTG-S5*	59.49	38.25	41.45	68.98	51.19	41.14	26.32	28.06	49.85	36.84

from 32.02%–33.33% to 33.98%–35.38%. Similar improvements are observed in mAP (from 33.20%–35.57% to 35.11%–35.88%), mAP@0.5 (from 63.65%–67.53% to 66.24%–67.61%), and mIoU (from 47.89%–48.66% to 49.06%–49.92%). On the TACoS dataset, R1@0.5 increases from 36.89%–38.42% to 38.32%–39.42%, R1@0.7 from 21.54%–22.59% to 22.57%–23.94%, and mAP from 20.95%–22.23% to 21.41%–23.05%. Similarly, consistent gains are observed in mAP@0.5 (from 41.24%–44.80% to 42.67%–44.23%), and mIoU (from 34.13%–36.75% to 35.72%–36.67%). These improvements across all random seeds and evaluation metrics further validate the effectiveness of FDAP in enhancing CG-DETR's performance.

To further evaluate the stability and robustness of FDAP across different random initializations, we report results of Flash-VTG variants under five distinct random seeds (S1–S5) in Table 11. Each row pair compares the baseline Flash-VTG with its FDAP-enhanced counterpart. Across both Charades-STA and TACoS datasets, we observe consistent performance gains after applying FDAP. In particular, FDAP contributes to improvements in stricter metrics such as R1@0.7, which reflect precise temporal grounding. For example, on Charades-STA, Flash-VTG-S4* achieves the highest R1@0.7 (38.39%), surpassing all other settings. Similarly, on TACoS, Flash-VTG-S4* yields the best R1@0.5 (41.04%). These consistent improvements across multiple runs validate the generalizability of FDAP, indicating that it not only enhances overall performance but also reduces variability across training seeds.

Overall, we conduct comprehensive experiments on multiple baseline models (M-DETR, TR-DETR, CG-DETR, Flash-VTG) across the QVHighlights, Charades-STA, and TACoS datasets, evaluating the effectiveness of our proposed FDAP module under five fixed random seeds. The results demonstrate that FDAP consistently improves retrieval and detection metrics across all models and datasets. Notably, FDAP yields substantial gains on stricter metrics such as R1@0.7, which require precise temporal grounding, confirming its ability to improve both accuracy and reliability. These findings validate FDAP as a generalizable and effective paradigm for temporal grounding models.