Multi-Scale Contrastive Learning for Video Temporal Grounding

Thong Thanh Nguyen[1], **Yi Bin**[2*], **Xiaobao Wu**[3], **Zhiyuan Hu**[1], **Cong-Duy T Nguyen**[3], **See-Kiong Ng**[1], **Anh Tuan Luu**[3]

[1] Institute of Data Science (IDS), National University of Singapore, Singapore
[2] Tongji University, China,
[3] Nanyang Technological University (NTU), Singapore

Abstract

Temporal grounding, which localizes video moments related to a natural language query, is a core problem of vision-language learning and video understanding. To encode video moments of varying lengths, recent methods employ a multi-level structure known as a feature pyramid. In this structure, lower levels concentrate on short-range video moments, while higher levels address long-range moments. Because higher levels experience downsampling to accommodate increasing moment length, their capacity to capture information is reduced and consequently leads to degraded information in moment representations. To resolve this problem, we propose a contrastive learning framework to capture salient semantics among video moments. Our key methodology is to leverage samples from the feature space emanating from multiple stages of the video encoder itself requiring neither data augmentation nor online memory banks to obtain positive and negative samples. To enable such an extension, we introduce a sampling process to draw multiple video moments corresponding to a common query. Subsequently, by utilizing these moments' representations across video encoder layers, we instantiate a novel form of multi-scale and cross-scale contrastive learning that links local short-range video moments with global long-range video moments. Extensive experiments demonstrate the effectiveness of our framework for not only long-form but also short-form video grounding.

Introduction

Temporal video grounding aims to localize moments of interest in an untrimmed video given a free-form textual description. It is a challenging multimodal task since it involves understanding temporal information in videos and reasoning about their connections to semantic information in texts. Recently, temporal grounding has drawn increasing attention (Mu, Mo, and Li 2024; Jung et al. 2023; Xu et al. 2023; Pan et al. 2023), due to its wide range of applications such as surveillance (Zhang, Zhu, and Roy-Chowdhury 2016), robotics (Burgner-Kahrs, Rucker, and Choset 2015), and autonomous driving (Claussmann et al. 2019).

Previous methods (Zhang et al. 2020b; Soldan et al. 2021; Zhang et al. 2020a) for temporal grounding concentrate on grounding merely a few queries in short video snippets. However, recently the growing availability of long videos, *e.g.* on streaming platforms, and demands to query their rich content have necessitated productive grounding of large volumes of queries in long videos. Because of such short-to-long video paradigm shift, latest methods (Zhang, Wu, and Li 2022; Mu, Mo, and Li 2024) have utilized local self-attention to restrict attention within a local window, following the intuition that temporal context beyond a certain range is less helpful for moment localization.

To capture moments at different temporal scales without enlarging the window size of the local self-attention, recent methods (Zhang, Wu, and Li 2022; Mu, Mo, and Li 2024) need to combine several Transformer blocks with downsampling between every two blocks, resulting in a feature pyramid of moment representations, as illustrated in Figure 2 (left). Unfortunately, due to such downsampling operation, when moment representations are propagated from lower levels of short-range (local) moments to higher levels of long-range (global) moments, information contained in representations of longer moments will gradually degrade (Guo et al. 2020; Yang et al. 2023). This could explain why performance of these methods tends to degrade as the duration of target moments increase, as shown in Figure 2 (right) and statistically shown with Intersection-over-Union (IoU) results in Figure 3, respectively.

To enrich information in video moment representations, recent works (Panta et al. 2024; Xiao et al. 2024; Ji et al. 2024; Liu et al. 2024) have employed contrastive learning for temporal grounding. The intuition is to capture mutual information between video moments and textual query to preserve salient semantics in moment representations. These works mainly involve query-moment pairs in which queries relate to video moments of distinct videos, hence the learned semantics among moment representations would be independent from each other. However, such approach might not be suitable for the latest scalable video-centric approach (Zhang, Wu, and Li 2022; Mu, Mo, and Li 2024), in which multiple textual queries are related to one video. Therefore, if the grounding of two textual queries results in temporal overlapping, there might be a conflict in compact moment representations (An et al. 2023). Furthermore, focusing upon moment-query relations limits these works to the feature space of the final encoder layer, which could not effectively utilize all hidden representations across encoding

*Yi Bin is the corresponding author, yi.bin@hotmail.com

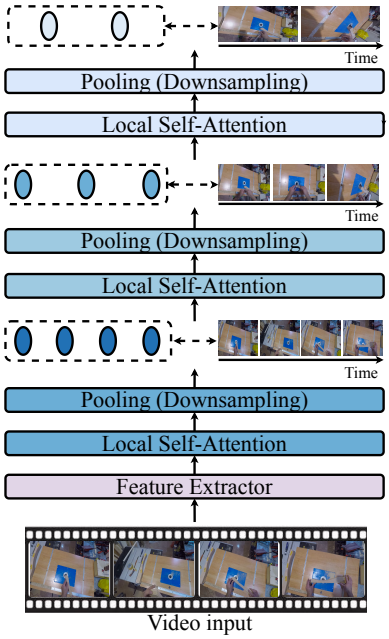

Figure 1: (Left) Illustration of feature pyramid to encode video moments of different lengths.

layers. For multi-scale temporal grounding, such cross-scale representations should be fully used since they express semantics in video moments of various lengths.

To resolve the above issues, in this paper, we propose a multi-scale contrastive learning framework for multi-scale temporal grounding. In our framework, instead of leveraging moment-query relationships, we utilize the association among video moments. Particularly, to avoid representation conflict among video moments, we introduce a query-centric contrastive approach that draws temporally separate video moments corresponding to a common textual query. A central component of our framework is the creation of positive and negative video moment samples, which previous works primarily apply data augmentation (Kim et al. 2022; Xing et al. 2023). However, because most long-form videos consist of a high volume of video moments, choosing an appropriate augmentation strategy that suits every moment is a non-trivial and lengthy tuning step. Another common approach is to introduce a memory bank to store positive or negative samples' representations, which are created by aggregating input representations iteratively during training (Panta et al. 2024; Han et al. 2023). Nevertheless, a memory bank would present additional hyperparameters such as the bank size and update frequency, which demand laborious tuning effort (Wang et al. 2021).

To prevent these problems, we directly draw samples from the feature space of video moment encoder. Specifically, we take advantage of internal, intermediate representations of video moments from the encoder that are readily available through the feed-forward step of the network without the need to rely upon external steps such as data augmentation or online storing of samples in memory banks. Accordingly, we introduce a within-scale and cross-scale approach to create positive and negative moment samples for contrastive learning. Regarding the within-scale approach, we seek to pull together representations of such semantically close video moments on the same scale of similar temporal range. Moreover, we also push apart representations of video moments which are unrelated to the textual query. Regarding the cross-scale approach, we compel the model to relate global long-range video moments to local short-range moments, while simultaneously repelling semantically distant cross-scale representations in an analogous cross-scale manner. This cross-scale approach would enable long-range moment representations to capture nuanced details of short-range moments, thereby mitigating informational degradation within long-range representations.

To sum up, our contributions are the following:

- We propose a multi-scale contrastive framework that focuses on moment-moment relations to mitigate informational degradation in video moment representations.
- We propose a within- and cross-scale strategy that supports semantic consistency not only between similar-range but also cross-range video moment representations emanating across layers of the video encoder.
- Our framework achieves superior results across major benchmark datasets concerning both short-form and long-form video grounding.

Methodology

In this section, we delineate our proposed contrastive framework for multi-scale temporal grounding, particularly focusing on a sampling procedure to draw video moment representations across temporal scales.

Preliminary - Video Temporal Grounding

We denote an input video V as a sequence of video clips $\{v_t\}_{t=1}^{T} = \{v_1, v_2, ..., v_T\}$, where v_t denotes a video moment (clip) centered at time t. We use a pre-trained feature extractor to embed each v_t into a moment embedding $\mathbf{v}_t$. Given the video V, our task is to localize a moment $y = (s, e)$ based on a sentence query $Q = \{q_1, q_2, ..., q_K\}$. Similar to the input video, we also embed the query Q into a sequence of word embeddings $\{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_K\}$.

Video encoder. After embedding video clips, we use a convolution-based projection function to encode local context of video clips:

$$Z^0 = \{\mathbf{z}_t^0\}_{t=1}^{T} = \text{Conv}\left(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T\right). \quad (1)$$

Subsequently, we designate L Transformer layers to encode temporal context among video clips. In detail, each Transformer layer consists of a local multi-head self-attention (LocalMSA) with a window size of W and an MLP block, in which we restrict the attention to be within a local window:

$$\bar{Z}^l = \alpha^l \cdot \text{LocalMSA}\left(\text{LN}\left(Z^{l-1}\right)\right) + Z^{l-1}, \quad (2)$$

$$\hat{Z}^l = \bar{\alpha}^l \cdot \text{MLP}\left(\text{LN}\left(\bar{Z}^l\right)\right) + \bar{Z}^l, \quad (3)$$

$$Z^l = \downarrow\left(\hat{Z}^l\right), \quad l \in \{1, 2, ..., L\}, \quad (4)$$

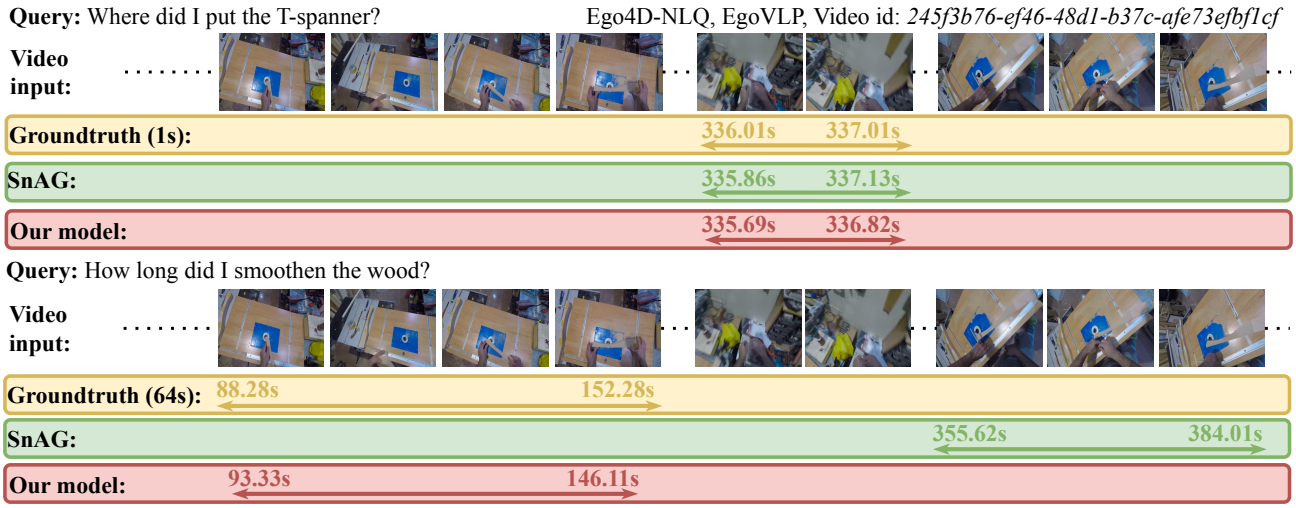

Figure 2: An example where recent method SnAG (Mu, Mo, and Li 2024) accurately localizes short video moment but fails on long moment.

where $Z^{l-1}, \bar{Z}^l, \hat{Z}^l \in \mathbb{R}^{T^{l-1} \times D}$, $Z^l \in \mathbb{R}^{T^l \times D}$. T^{l-1}/T^l is the downsampling ratio, α^l and $\bar{\alpha}^l$ are learnable per-channel scaling factors (Touvron et al. 2021), D is the hidden dimension, and LN is the layer normalization.

Inspired by (Mu, Mo, and Li 2024), we implement the downsampling operator $\downarrow$ as a strided depthwise 1D convolution. The downsampling operation engenders the multi-scale property of the encoder, generating representations for longer video moments.

Text encoder. We use Transformer layers, where each layer includes a vanilla self-attention followed by an MLP. Thus, the textual encoder produces textual representations $E = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K\}$ for query embeddings $\{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_K\}$.

Cross-modal fusion. Our architecture uses cross-attention to fuse video clip and query word representations. Technically, we modulate video clip representations $\{Z^l\}_{l=1}^L$ with word representations E as follows:

$$\tilde{Z}^l = \text{LN}\left(Z^l\right), \quad \tilde{E} = \text{LN}\left(E\right), \tag{5}$$

$$O^l = \sigma\left(\frac{\left(\tilde{Z}^l\right)^\top \cdot \tilde{E}}{\sqrt{D}}\right) \cdot \tilde{Z}^l, \tag{6}$$

$$X^l = \beta^l \cdot \text{MLP}\left(\text{LN}\left(O^l\right)\right) + O^l, \tag{7}$$

where β^l denotes a learnable per-channel scale and σ the Softmax activation function.

Moment decoding. After cross-modal fusion, our model converts each time step t to a moment candidate. Specifically, given $\mathbf{x}_t^l$, we use a convolutional network comprising 1D convolutional layers as the classification head to predict a score p_t^l. In a similar vein, we use a similar 1D convolutional network attached with a ReLU activation function to regress the normalized distances from t to the moment boundaries (d_t^s, d_t^e) if $\mathbf{x}_t^l$ is classified as positive. Formally, the decoded moment is computed as:

$$(t, l) = \arg\max_{t,l} p_t^l, \tag{8}$$

$$\hat{s} = 2^{l-1}\left(t - d_t^s\right), \quad \hat{e} = 2^{l-1}\left(t + d_t^e\right). \tag{9}$$

During testing, we employ Soft-NMS (Bodla et al. 2017) to merge overlapping moment predictions.

Cross-scale Contrastive Learning

Query-centric sampling. As randomly sampling moment-query pairs for contrastive learning might lead the model to representation conflict if the groundings of two queries overlap with each other, we instead introduce a sampling approach that draws a text query Q and its temporally separate video moments associated with a common video V:

$$Q_{j'}, \{y_{j'}^l\}_{l=1}^L \sim \mathcal{U}\left(\{Q_j, \{y_{j'}^l\}_{l=1}^L\}_{j=1}^{N_Q}\right), \tag{10}$$

where $\mathcal{U}$ denotes a discrete uniform distribution, $\{y_{j'}^l\}_{l=1}^L$ the set of target video moments in each layer l, and N_Q the number of textual queries related to video V. We generate the target set $\mathcal{P}(l)$ via center sampling (Zhang, Wu, and Li 2022; Mu, Mo, and Li 2024), *i.e.* given any moment centered at t, any time step $c \in \left[t - \alpha\frac{T}{T^l}, t + \alpha\frac{T}{T^l}\alpha\right]$ in layer l is considered as a target. After sampling the query and target moments, we directly utilize the representations $\{\mathbf{z}_{j'}^l\}_{l=1}^L$ of the target moments $\{y_{j'}^l\}_{l=1}^L$ extracted by the aforementioned multi-scale video encoder.

Within-scale contrastive learning. Having obtained the representations of target moment samples, we directly utilize moments within each scale as positive and negative samples. Particularly, we iterate over every layer l of the video encoder, and for each anchor video moment $y_{j'}^l$, we consider all video moments of layer l corresponding to query $Q_{j'}$ to become positive moment set $P(l)$, and randomly draw those not corresponding to query $Q_{j'}$ to be negative set $\mathcal{N}(l)$.

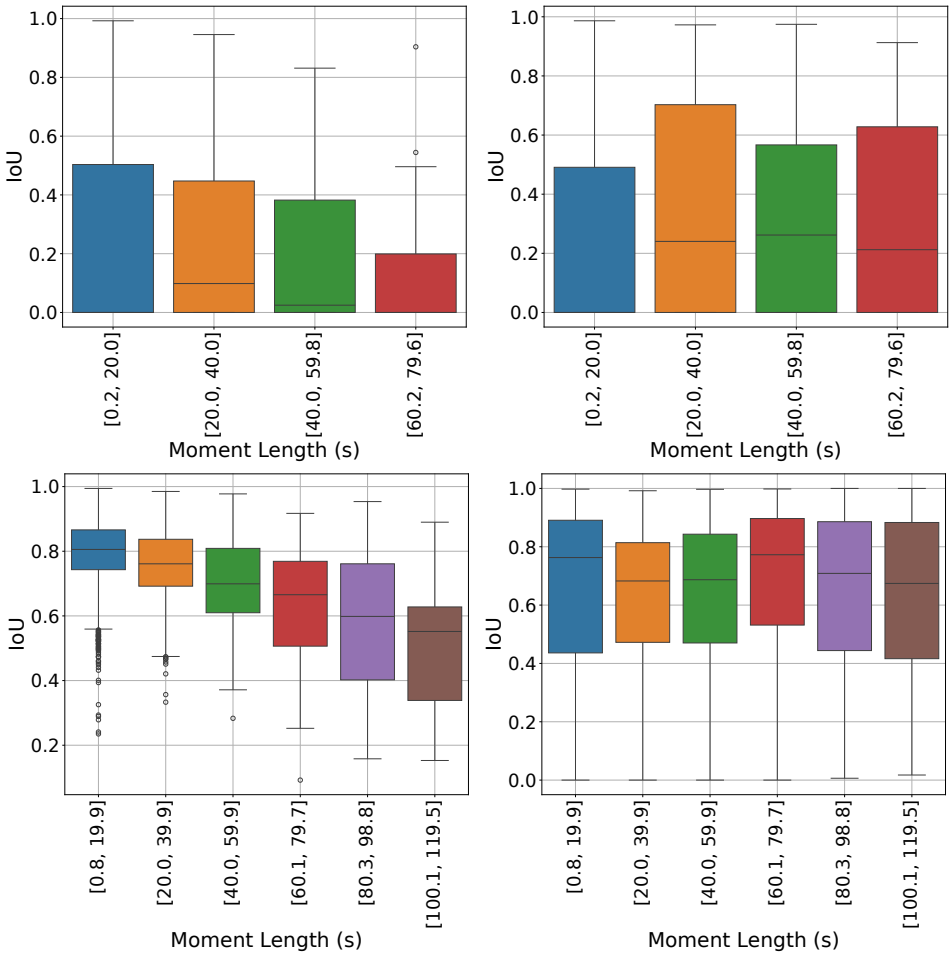

Figure 3: First and Second: IoU results with respect to target video moment length on Ego4D-NLQ (Grauman et al. 2022) of baseline SnAG (Mu, Mo, and Li 2024) and our model. Third and Fourth: IoU results with respect to target video moment length on TACoS (Regneri et al. 2013) datasets of baseline SnAG (Mu, Mo, and Li 2024) and our model.

Then, we formulate multi-scale contrastive objective over all layers $l \in \{1, 2, ..., L\}$, which pushes positive moments closer while negative moments further:

$$\mathcal{L}_{\text{within}} = -\sum_{l=1}^{L} \sum_{i \in \mathcal{P}(l)} \sum_{j \in \mathcal{P}(l), i \neq j} \log \frac{e^{\left(\mathbf{z}_i^l \cdot \mathbf{z}_j^l\right)}}{e^{\left(\mathbf{z}_i^l \cdot \mathbf{z}_j^l\right)} + \sum_{n \in \mathcal{N}(l)} e^{\left(\dot{\mathbf{z}}_i^l \cdot \mathbf{z}_n^l\right)}}. \tag{11}$$

Cross-scale contrastive learning. We further associate semantically close moment representations from across different scales. Specifically, we push short-range moment representations closer to semantically close long-range moment representations. This would enable short-range moments to relate to longer video context while long-range features to capture nuanced details of short-range moments.

As video moment features of layer 0 $\{\mathbf{z}_{j'}^0\}$ are the most likely to preserve salient video information compared to other levels, we employ features of the target moments from the lowest level as the anchor set for cross-scale contrastive learning. To construct positive and negative moment set, we utilize features of higher levels $l \in \{1, 2, ..., L\}$ in the feature pyramid corresponding to video moments that involve and do not involve the textual query, respectively. Denoting the set of moment indices in level l that are related to the query as $\mathcal{P}(l)$ and the set of those that are unrelated as $\mathcal{N}(l)$, we define the cross-scale contrastive learning objective as:

$$\mathcal{L}_{\text{cross}} = -\sum_{i \in \mathcal{P}(0)} \sum_{l=1}^{L} \sum_{j \in \mathcal{P}(l)} \log \frac{e^{\left(\mathbf{z}_i^0 \cdot \mathbf{z}_j^l\right)}}{e^{\left(\mathbf{z}_i^0 \cdot \mathbf{z}_j^l\right)} + \sum_{n \in \mathcal{N}(l)} e^{\left(\mathbf{z}_i^0 \cdot \mathbf{z}_n^l\right)}}. \tag{12}$$

Training Objective

For temporal grounding training, we adopt a focal loss $\mathcal{L}_{\text{cls}}$ for target moment classification and a Distance-IoU loss $\mathcal{L}_{\text{reg}}$ for distance regression from a positive time step t to the tar-

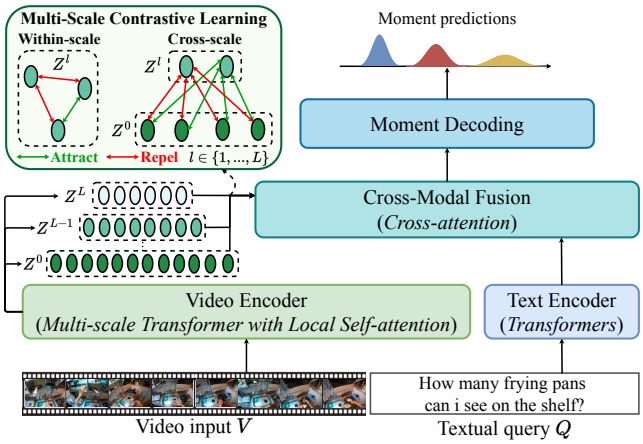

Figure 4: Overall illustration of the proposed framework.

get moment. Then, we combine these losses with our within- and cross-scale contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \rho_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \rho_{\text{within}} \cdot \mathcal{L}_{\text{within}} + \rho_{\text{cross}} \cdot \mathcal{L}_{\text{cross}}, \quad (13)$$

where ρ_{reg}, ρ_{within}, and ρ_{cross} denote hyperparameters to balance the regression, within-scale, and cross-scale contrastive losses, respectively.

Experiments

To validate the effectiveness, we conduct extensive experiments against recent methods for temporal grounding. We also perform ablation study to investigate each component.

Datasets

Following previous works, we work on five challenging datasets of temporal grounding, which belong to two main categories, *i.e.* 1) Long videos, many queries (Ego4D-NLQ (Grauman et al. 2022), MAD (Soldan et al. 2022), and TACoS (Regneri et al. 2013)) and 2) Short videos, few queries (ActivityNet-Captions (Krishna et al. 2017) and Charades-STA (Sigurdsson et al. 2016)).

Ego4D-NLQ (Grauman et al. 2022) consists of egocentric videos recording daily human activities. Each video possesses length from 3.5 to 20 minutes and is associated with 11.6 queries on average.

MAD (Soldan et al. 2022) comprises 1.2K hours of movies with 384K queries transcribed from audio description. Since each video is a movie, each exhibits 47 to 202 minutes long.

TACoS (Regneri et al. 2013) focuses on cooking topics. The total video length is 10.1 hours and each video is tasked with 143.5 queries for the temporal grounding operation.

ActivityNet-Captions (Krishna et al. 2017) targets dense video captioning and is subsequently adapted to temporal grounding. Its video length is two minutes on average and the average number of queries per video is approximately 3.65 queries.

Charades-STA (Sigurdsson et al. 2016) is an action recognition dataset transformed into a temporal grounding one. Each video lasts approximately 30 seconds and possesses 2.4 queries.

Evaluation Metrics

We report Recall@K at different temporal intersection-over-union θ (R@K, tIoU = θ) for all datasets. The metric measures the percentage of textual queries whose at least one of the top-K moment predictions temporally overlap with the groundtruth moment more than θ.

Implementation Details

To fairly compare with previous works and satisfy the scalability of temporal grounding operation for long videos, we adopt video-centric sampling approach (Mu, Mo, and Li 2024). For Ego4D-NLQ, we use pre-trained 1) SlowFast video features (Feichtenhofer et al. 2019) with BERT textual features (Devlin et al. 2018), and 2) EgoVLP video and textual features (Lin et al. 2022). For testing, we report R@$\{1, 5\}$, tIoU = $\{0.3, 0.5\}$. For MAD dataset, we use CLIP features (Radford et al. 2021) for both videos and texts, and report R@$\{1, 5, 10, 50\}$, tIoU = $\{0.1, 0.3, 0.5\}$. For the TACoS dataset, we use C3D video features (Tran et al. 2015) and GloVe textual features (Pennington, Socher, and Manning 2014). We report results in terms of R@$\{1, 5\}$, tIoU = $\{0.5, 0.7\}$. In addition, we utilize I3D features (Carreira and Zisserman 2017) pre-trained on Kinetics (Kay et al. 2017) for Charades-STA and C3D features (Tran et al. 2015) for ActivityNet-Captions experiments. For both datasets, similar to TACoS, we take advantage of GloVe textual features (Pennington, Socher, and Manning 2014). We report R@$\{1, 5\}$, tIoU = $\{0.5, 0.7\}$ for testing on Charades-STA, and R@$\{1, 5\}$, tIoU = $\{0.3, 0.5\}$ for testing on ActivityNet-Captions. For more details regarding model architecture, we direct interested readers to the appendix. For both within-scale and cross-scale contrastive learning implementation, we keep the size of the negative sample set $\mathcal{N}(l)$ in every level l to be equal to the size of the positive video clips $\mathcal{P}(l)$ that correspond to the target video moments. Based upon validation and fair comparison with previous methods, we use $\rho_{\text{ref}} = \rho_{\text{within}} = \rho_{\text{cross}}$ = 1.0.

Baselines

We consider the following temporal grounding models as baselines: (i) **VSL-Net** (Zhang et al. 2020a) utilizing textual query to highlight regions potential to comprise the target moment; (ii) **VLG-Net** (Soldan et al. 2021) modeling temporal grounding as a graph matching problem; (iii) **Moment-DETR** (Lei, Berg, and Bansal 2021), a Transformer encoder-decoder architecture that views temporal grounding as a set prediction problem; (iv) **CONE** (Hou et al. 2022) subsequently slicing a video input into windows, selects relevant windows, and ranks the selected windows to obtain target moments; (v) **MMN** (Wang et al. 2022), a Siamese-like network architecture that is trained with video-query and query-video contrastive learning; (vi) **SSRN** (Zhu et al. 2023) enriching anchor frames with additional consecutive frames; (vii) **G2L** (Li et al. 2023) measuring moment-query similarities using geodesic distance and quantifies cross-modal interactions with game-theoretic interactions; (viii) **SOONet** (Pan et al. 2023), an anchor-based framework that conducts grounding by pre-ranking,

Features	Model	R@1			R@5		
		0.3	0.5	Avg	0.3	0.5	Avg
SF+BERT	VSL-Net	5.45	3.12	4.29	10.74	6.63	8.69
	CONE	10.40	5.03	7.72	22.74	11.87	17.31
	SOONet	8.00	3.76	5.88	22.40	11.09	16.75
	SnAG	9.83	6.83	8.33	27.93	19.27	23.60
	Our model	**10.80**	**7.22**	**9.49**	**28.54**	**20.38**	**25.06**
EgoVLP	VSL-Net	10.84	6.81	8.83	18.84	13.45	16.15
	CONE	14.15	8.18	11.17	30.33	18.02	24.18
	SnAG	15.72	10.78	13.25	38.39	27.44	32.92
	Ours	**16.37**	**11.27**	**13.96**	**39.97**	**28.70**	**34.43**

Table 1: Results on Ego4D-NLQ.

re-ranking, and regression; (ix) **MESM** (Liu et al. 2024), a fine-grained moment-query contrastive approach modeled for query word and video moment representations; (x) **Contrastive-MSAT** (Panta et al. 2024), applying moment-query contrastive loss supported by a momentum-based memory bank; (xi) **UVCOM** (Xiao et al. 2024), a moment-query contrastive approach for a unified video comprehension framework; (xii) **SnAG** (Mu, Mo, and Li 2024) achieving scalable grounding with cross-modal late fusion.

Experimental Results

Main Results

Results on Ego4D-NLQ (Table 1). Our framework significantly outperforms recent temporal grounding methods. For example, using SlowFast+BERT features, we outperform previous best method, *i.e.* SnAG, by mean improvements of 1.16% and 1.46% in terms of R@1 and R@5 metrics, respectively. In addition, we accomplish more significant performance gains on the more stringent tIoU threshold of 0.5, denoting more precise moment localization.

Results on MAD (Table 2). Similar to results on Ego4D-NLQ, our framework obtains an outstanding improvement over previous temporal grounding methods. Specifically, we enhance SOONet with 1.68 and 2.82 points of R@1 and R@5 on average. Moreover, our model outperforms CONE and SnAG in terms of mean R@1 / R@5 by 3.58 / 6.08 and 2.11 / 1.90 points, respectively, especially for the more stringent tIoU threshold.

Results on TACoS (Table 3 (left)). Our model achieves R@1 / R@5 of 47.04% / 73.55% at tIoU = 0.5, outperforming the strongest baseline, *i.e.* SnAG, by a substantial margin, *i.e.* +2.18% R@1 and +2.89% R@5. Combined with the results on Ego4D-NLQ and MAD, these results demonstrate that our contrastive framework provides beneficial signals to counter informational degradation in the feature pyramid for long-form video grounding.

Results on ActivityNet-Captions (Table 3 (middle)). We achieve R@1 / R@5 scores of 33.56% / 68.91% at tIoU = 0.7. These results indicate that we outperform SSRN by 0.41% and 0.43% with regards to R@1 and R@5, respectively, even though we use the backbone SnAG which is significantly weaker than SSRN.

Results on Charades-STA (Table 3 (right)). Our model outperforms previous methods by a wide margin. Particularly, we accomplish 47.03% R@1 and 72.53% R@5 at tIoU = 0.7, exceeding SSRN by 4.38% R@1 and 7.04% R@5. These outcomes on Charades-STA and ActivityNet-Captions show that mutual information signals among video moments contributed by our contrastive framework can polish video moment representations to help temporal grounding on short-form videos.

Ablation Study

We conduct extensive experiments on TACoS to study the influence of the design choices.

Effect of contrastive components. We explore what extent each component of our contrastive framework, *i.e.* within- or cross-scale objective, contributes to the overall performance improvement. In Table 5, cross-scale objective plays a more fundamental role in polishing video moment representations than the within-scale counterpart. Since cross-scale contrastive objective concentrates more upon long-range moment representations by relating them with the short-range moment ones, these results validate our hypothesis that informational degradation is a fundamental problem to resolve in multi-scale temporal grounding.

Effect of moment-moment association. In addition to our proposed moment-moment association, we experiment with various approaches, *i.e.* moment-query association, query-query association, and one approach to associate video moments but based on the semantic closeness of their corresponding textual queries. For the last approach, we consider two textual queries to be semantically similar if their CLIP-based cosine similarity score is greater than or equal to 0.8 (for positive sampling) and semantically distant if the similarity score is smaller than or equal to 0.2 (for negative sampling). As can be observed in Table 6, query-query association performs the worst, as the approach does not polish moment representations. The moment-moment approach outperforms moment-query contrastive learning, but underperforms our method. We hypothesize that there might exist representation conflict between two video moments temporally overlap with each other.

Effect of direct utilization of moment representations. We study the impact of our direct utilization of moment representations for positive and negative sample generation, and compare with Tube TokenMix (Xing et al. 2023) as the data augmentation and the momentum-based memory bank approach (Panta et al. 2024). Table 4 shows that we significantly surpass other methods, on average by 1.38 / 1.60 points of R@1 / R@5 over the augmentation approach, and 0.45 / 0.66 points of R@1 / R@5 over the memory bank approach. We hypothesize that while memory bank may maintain a high number of samples for contrastive learning, expensive hyperparameter tuning is essential to achieve an effective performance.

Qualitative Analysis

In Figure 3, we observe that our model does not encounter degraded performance when the lengths of the target moments increase. Moreover, we visualize moment predictions of the recent method, *i.e.* SnAG (Mu, Mo, and Li 2024), and

Model	R@1			R@5			R@10			R@50		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
VLG-Net	3.64	2.76	1.65	11.66	9.31	5.99	17.39	14.56	9.77	39.78	34.27	24.93
Moment-DETR	0.31	0.24	0.16	1.52	1.14	0.28	2.79	2.06	1.20	11.08	7.97	4.71
CONE	8.90	6.87	4.10	20.51	16.11	9.59	27.20	21.53	12.82	43.36	34.73	20.56
SOONet	11.26	9.00	5.32	23.21	19.64	13.14	30.36	26.00	17.84	50.32	44.78	32.59
SnAG	10.28	8.46	5.55	24.42	20.60	13.75	32.23	27.50	19.00	52.28	46.68	35.24
Our model	**12.76**	**10.94**	**6.92**	**26.43**	**22.60**	**15.43**	**34.08**	**29.41**	**20.70**	**54.84**	**48.26**	**37.77**

Table 2: Results on MAD.

Model	TACoS				ActivityNet-Captions				Charades-STA			
	R@1		R@5		R@1		R@5		R@1		R@5	
	0.3	0.5	0.3	0.5	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7
VLG-NET	45.46	34.19	70.38	56.56	46.32	29.82	77.15	63.33	-	-	-	-
MGSL-Net	42.54	32.27	63.39	50.13	51.87	31.42	82.60	66.71	63.98	41.03	93.21	63.85
MMN	39.24	26.17	62.03	47.39	48.59	29.26	79.50	64.76	-	-	-	-
SSRN	45.10	34.33	65.26	51.85	54.49	33.15	84.72	68.48	65.59	42.65	94.76	65.48
G2L	42.74	30.95	65.83	49.86	51.68	33.35	81.32	67.60	-	-	-	-
MESM	52.69	39.52	-	-	-	-	-	-	61.24	38.04	-	-
Contrastive-MSAT	49.77	37.99	68.31	58.31	47.73	31.21	78.06	63.63	-	-	-	-
UVCOM	36.39	23.32	-	-	-	-	-	-	59.25	36.64	-	-
SnAG	56.44	44.86	81.15	70.66	48.55	30.56	81.71	63.41	64.62	46.26	92.55	71.94
Ours	**58.17**	**47.04**	**84.84**	**73.55**	**54.83**	**33.56**	**84.78**	**68.91**	**66.64**	**47.03**	**93.66**	**72.53**

Table 3: Results on TACoS, ActivityNet-Captions, and Charades-STA.

Positive-negative sampling approach	R@1		R@5	
	0.3	0.5	0.3	0.5
Data augmentation	57.00	45.46	83.13	72.06
Memory bank	57.69	46.62	84.13	72.94
Ours	**58.17**	**47.04**	**84.84**	**73.55**

Table 4: Ablation results on TACoS with various positive and negative sampling approaches.

Contrastive component	R@1		R@5	
	0.3	0.5	0.3	0.5
w/o within-scale	57.40	46.00	83.46	72.39
w/o cross-scale	57.00	45.85	82.34	71.58
Ours	**58.17**	**47.04**	**84.84**	**73.55**

Table 5: Ablation results on TACoS with multi-scale contrative components.

our model in Figure 2. Even though SnAG could precisely detect the shorter-length moment, it misses the moment of longer length, due to the degraded information issue. In contrast, our framework is able to localize both the short and long moments. We hypothesize that our contrastive framework can hold salient semantics for video moment representations to resolve the degraded signals in the grounding model, thus enhancing the grounding operation towards long video moments.

Association approach	R@1		R@5	
	0.3	0.5	0.3	0.5
Query-query	55.61	45.06	81.25	71.75
Moment-query	57.00	46.24	82.44	72.37
CLIP-based moment-moment	57.13	46.94	83.28	72.96
Ours	**58.17**	**47.04**	**84.84**	**73.55**

Table 6: Ablation results on TACoS with various association approaches.

Conclusion

In this paper, we propose a multi-scale contrastive framework for multi-scale temporal grounding. Essentially, our framework utilizes a query-centric approach to associate temporally separate video moments which correspond to a common textual query to avoid representation conflict. Accordingly, we define a within-scale contrastive objective to model relations among similar-range video moments, and a cross-scsale objective to model relations among cross-range moments. Comprehensive experiments validate the effectiveness of our framework for both short-term and long-term temporal grounding.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-051T). Thong Nguyen is supported by a Google Ph.D. Fellowship in Natural Language Processing.

References

An, X.; Deng, J.; Yang, K.; Li, J.; Feng, Z.; Guo, J.; Yang, J.; and Liu, T. 2023. Unicom: Universal and compact representation learning for image retrieval. *arXiv preprint arXiv:2304.05884*.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Burgner-Kahrs, J.; Rucker, D. C.; and Choset, H. 2015. Continuum robots for medical applications: A survey. *IEEE Transactions on Robotics*, 31(6): 1261–1280.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Claussmann, L.; Revilloud, M.; Gruyer, D.; and Glaser, S. 2019. A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(5): 1826–1848.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; and Pan, C. 2020. Augfpn: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12595–12604.

Han, D.; Cheng, X.; Guo, N.; Ye, X.; Rainer, B.; and Priller, P. 2023. Momentum cross-modal contrastive learning for video moment retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hou, Z.; Zhong, W.; Ji, L.; Gao, D.; Yan, K.; Chan, W.-K.; Ngo, C.-W.; Shou, Z.; and Duan, N. 2022. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*.

Ji, W.; Shi, R.; Wei, Y.; Zhao, S.; and Zimmermann, R. 2024. Weakly Supervised Video Moment Retrieval via Location-irrelevant Proposal Learning. In *Companion Proceedings of the ACM on Web Conference 2024*, 1595–1603.

Jung, M.; Jang, Y.; Choi, S.; Kim, J.; Kim, J.-H.; and Zhang, B.-T. 2023. Overcoming Weak Visual-Textual Alignment for Video Moment Retrieval. *arXiv preprint arXiv:2306.02728*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, T.; Kim, J.; Shim, M.; Yun, S.; Kang, M.; Wee, D.; and Lee, S. 2022. Exploring temporally dynamic data augmentation for video recognition. *arXiv preprint arXiv:2206.15015*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2023. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12032–12042.

Lin, K. Q.; Wang, J.; Soldan, M.; Wray, M.; Yan, R.; Xu, E. Z.; Gao, D.; Tu, R.-C.; Zhao, W.; Kong, W.; et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35: 7575–7586.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3855–3863.

Mu, F.; Mo, S.; and Li, Y. 2024. SnAG: Scalable and Accurate Video Grounding. *arXiv preprint arXiv:2404.02257*.

Pan, Y.; He, X.; Gong, B.; Lv, Y.; Shen, Y.; Peng, Y.; and Zhao, D. 2023. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13767–13777.

Panta, L.; Shrestha, P.; Sapkota, B.; Bhattarai, A.; Manandhar, S.; and Sah, A. K. 2024. Cross-modal Contrastive Learning with Asymmetric Co-attention Network for Video Moment Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 607–614.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

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

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 510–526. Springer.

Soldan, M.; Pardo, A.; Alcázar, J. L.; Caba, F.; Zhao, C.; Giancola, S.; and Ghanem, B. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5026–5035.

Soldan, M.; Xu, M.; Qu, S.; Tegner, J.; and Ghanem, B. 2021. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3224–3234.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; and Van Gool, L. 2021. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7303–7313.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.

Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18709–18719.

Xing, Z.; Dai, Q.; Hu, H.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2023. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18816–18826.

Xu, M.; Soldan, M.; Gao, J.; Liu, S.; Pérez-Rúa, J.-M.; and Ghanem, B. 2023. Boundary-denoising for video activity localization. *arXiv preprint arXiv:2304.02934*.

Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; and Liang, R. 2023. AFPN: asymptotic feature pyramid network for object detection. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2184–2189. IEEE.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

Zhang, S.; Zhu, Y.; and Roy-Chowdhury, A. K. 2016. Context-aware surveillance video summarization. *IEEE Transactions on Image Processing*, 25(11): 5469–5478.

Zhu, J.; Liu, D.; Zhou, P.; Di, X.; Cheng, Y.; Yang, S.; Xu, W.; Xu, Z.; Wan, Y.; Sun, L.; et al. 2023. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*.