SHINE: Saliency-aware HIerarchical NEgative Ranking for Compositional Temporal Grounding

Zixu Cheng*[1], Yujiang Pu*[2], Shaogang Gong[1], Parisa Kordjamshidi[2], and Yu Kong[2]

[1] Queen Mary University of London {zixu.cheng, s.gong}@qmul.ac.uk
[2] Michigan State University {puyujian, kordjams, yukong}@msu.edu

Abstract. Temporal grounding, a.k.a video moment retrieval, aims at locating video segments corresponding to a given query sentence. The compositional nature of natural language enables the localization beyond predefined events, posing a certain challenge to the compositional generalizability of existing methods. Recent studies establish the correspondence between videos and queries through a decompose-reconstruct manner to achieve compositional generalization. However, they only consider dominant primitives and build negative queries through random sampling and recombination, resulting in semantically implausible negatives that hinder the models from learning rational compositions. In addition, recent DETR-based methods still underperform in compositional temporal grounding, showing irrational saliency responses when given negative queries that have subtle differences from positive queries. To address these limitations, we first propose a large language model-driven method for negative query construction, utilizing GPT-3.5-Turbo to generate semantically plausible hard negative queries. Subsequently, we introduce a coarse-to-fine saliency ranking strategy, which encourages the model to learn the multi-granularity semantic relationships between videos and hierarchical negative queries to boost compositional generalization. Extensive experiments on two challenging benchmarks validate the effectiveness and generalizability of our proposed method. Our code is available at https://github.com/zxccade/SHINE.

Keywords: Temporal Grounding · Compositional Generalization

1 Introduction

Temporal grounding [8, 17, 18, 29, 50, 52] in videos has received continuous attention for its wide range of applications in moment retrieval and automatic content generation. Unlike typical temporal action localization tasks [54], temporal grounding aims to retrieve video moments based on textual queries that include not only the action itself but also the objects, attributes, and interactions involved. Moreover, the extensive range of vocabulary in natural language can

* Equal Contribution.

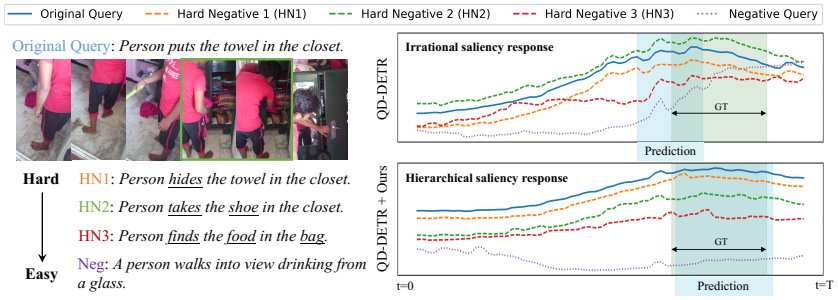


Fig. 1: Comparison of saliency scores given different queries. The existing work [29] struggles with discerning hard negative queries, showing irrational saliency responses under different primitive substitutions. Our method helps a model to learn the nuances in the semantics of hierarchical negative samples, suppressing the model's response to irrelevant queries while boosting its compositional generalizability.

expand the limited words from labeled videos to describe a variety of new things and scenarios unlabeled in training data. However, recent works [5,33,34,47] have shown that existing vision-language models (VLMs) lack compositional generalizability for unseen combinations, as reflected by their insensitivity to word order and significant primitives.

To address this challenge, compositional temporal grounding has been recently proposed in [20], which aims at locating unseen video moments by learning novel combinations of known words in the training data. They decompose videos and queries into global events, local actions, and atomic objects, and establish relationships between visual and text concepts by constructing a hierarchical semantic graph. However, their pipeline greatly relies on off-the-shelf object detection and action recognition models, which exhibits poor flexibility and scalability. Later, Li *et al*. [19] proposed a self-supervised learning framework to enhance the compositional generalization capability of existing VLMs by masking different primitives to generate semantically equivariant and invariant samples. In contrast, Deco [42] has constructed negative samples through a decompose-reconstruct strategy, employing a mask-and-predict ranking loss to learn the multi-granularity correspondence between video-text pairs.

While these methods have improved existing techniques, they share limitations in the construction of negative samples: (1) VISA [20] and SSL2CG [19] focus on dominant verbs and nouns, but overlook the role of other primitives such as prepositions and adverbs (*e.g.*, *on/under the table* and *turn on/off the light*), where the substitution of these words fundamentally changes the semantics. (2) Deco [42] neglects the viability of semantics when recombining primitives, resulting in numerous infeasible combinations, such as *eating the table* and *reading the door*. These issues not only hinder the model from learning the semantics of non-dominant primitives, but also force the model to extract differences from unrealistic combinations.

Additionally, these works [19, 20, 42] have only explored the compositional generalizability of classic temporal grounding methods, lacking considerations

on recent novel architectures, such as DETR-based methods [14, 18, 25, 29]. These approaches combine highlight detection [43] with the temporal grounding task, aiming to locate segments corresponding to the query while predicting the saliency scores for each moment. The saliency scores evaluate the relevance of all video clips to a given query, revealing the corresponding highlight moments. However, we observed that existing work [29] struggles with discerning different negative queries, showing irrational saliency responses under different primitive substitutions, as shown in Fig. 1. This indicates that current approaches tend to ignore the nuances between hard negative and positive queries. Consequently, they fail to accurately match visual representations with corresponding primitives, hindering their ability to achieve compositional generalization.

To this end, we first propose a large language model (LLM)-driven approach for constructing hard negative samples, which are semantically plausible yet distinct from the original query. With these manipulated negatives at hand, we further introduce a coarse-to-fine saliency ranking strategy to establish a multi-granularity semantic relationship between video clips and hierarchical negative queries. Compared to existing works, our method has the following advantages: (1) a good compositional representation for negative queries that consider the significance of different primitives while maintaining semantic feasibility; (2) the saliency scores derived from negative samples at different levels exhibit a hierarchical divergence, indicating that our method successfully captures the multi-granularity relationship between video clips and queries; (3) our method can be seamlessly integrated into existing DETR-based models, significantly improving their generalization capabilities to unseen combinations while maintaining the accuracy for seen samples. In summary, our contributions are three-fold:

- To address the issue of implausible negative queries generated by random sampling, we introduce an LLM-driven approach that produces semantically viable hard negative queries, which facilitates temporal grounding models to learn plausible compositional semantics.
- To deal with the irrational saliency responses in existing methods, we propose a coarse-to-fine saliency ranking strategy that utilizes the plausible hard negatives to capture hierarchical semantic differences and boost their compositional generalizability.
- Extensive experiments are conducted with two DETR-based backbones on two challenging benchmarks, Charades-CG and ActivityNet-CG, which show that our method significantly improves baseline performance and achieves competitive results.

2 Related Work

Temporal Grounding. Temporal grounding, *a.k.a.* video moment retrieval, initially proposed in [8] and [17], aims at localizing segments in a video that match the description of a query sentence. Currently, dominant supervised learning techniques are divided into two categories: proposal-based and proposal-free methods. Proposal-based methods generate candidate segments through various

strategies, including sliding windows [8,24], dense proposals [12,37,40], and fixed anchors [4, 35, 45, 49, 52], subsequently selecting the most appropriate intervals based on a similarity measure. However, the generation of candidate proposals and their semantic matching with queries are computationally intensive. Conversely, proposal-free methods [11,22,30,41,46,48] directly predict the temporal boundaries of the target clip. This paradigm eliminates the need for proposal generation, significantly enhancing the model's efficiency during inference. Recently, Lei *et al*. [18] reformulated the temporal grounding task as a set prediction problem, introducing a DETR-based [2] architecture enabling simultaneous video moment retrieval and highlight detection. Subsequent works, including UMT [25], QD-DETR [29], and EaTR [14], have enhanced localization accuracy by refining the DETR framework. Differently, our work establishes a connection between the saliency score and compositionality, which effectively unlocks the potential for compositional generalization in DETR-based models.

Compositional Generalization. Recently, the compositional generalizability of vision-language models, or VLMs, has received sustained attention [5,34,38,39, 47], with several benchmarks being proposed for evaluating the robustness of the models on specific downstream tasks, including image-text retrieval [13, 27, 32], visual question answering [7, 10, 15, 44], and zero-shot learning [21, 26, 53]. To further evaluate the compositional generalization of existing temporal grounding methods, Li *et al*. [20] has constructed two benchmarks, Charades-CG and ActivityNet-CG, and proposed a variational cross-graph reasoning framework to achieve compositional video-text comprehension. Later, Li *et al*. [19] generated semantic equivariant and invariant samples by masking different primitives, and employed contrastive learning to improve the compositional generalization capability of existing VLMs. Yang *et al*. [42] constructed negative queries through a decompose-reconstruct strategy, utilizing a mask-and-predict contrastive ranking loss to learn the multi-granularity correspondence between video-text pairs. However, most of these works only consider dominant primitives like verbs and nouns, ignoring the effect of other words like prepositions and adverbs. Moreover, they adopt random sampling to replace primitives for negative construction, which hinders the model from learning semantically feasible compositions. In contrast, we progressively replace the primitives with different ratios beyond just verbs and nouns. Additionally, we resort to a large language model to generate semantically plausible negative queries instead of random sampling.

3 Method

3.1 Problem Definition and Overview

Given an untrimmed video V and a query sentence Q, our goal is to identify the start and end timestamps (t_s, t_e) of the moments in the video that correspond to the query. The model is expected to achieve precise localization based on the novel combinations of seen words in the training set. Compared to the conventional temporal grounding task, we seek a good balance in performance between

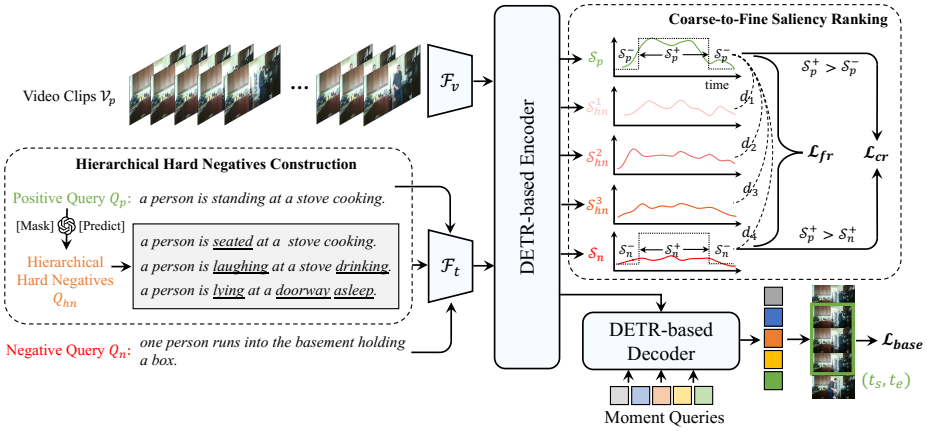


Fig. 2: The overall framework of our method SHINE. For each video-text pair, we first generate a set of hierarchical hard negative queries and randomly sample one negative query from the same mini-batch. These queries and the video clips are fed into a DETR-based encoder for interaction and predicting saliency scores S. The coarse-grained ranking loss $\mathcal{L}_{cr}$ aims to enlarge the disparity between the saliency scores produced by positive and negative queries, and the fine-grained ranking loss $\mathcal{L}_{fr}$ is designed to capture the nuanced semantics among the hierarchical hard negative queries. These two constraints are combined with $\mathcal{L}_{base}$ to optimize the model.

seen and unseen compositions, which requires the model to avoid overfitting while ensuring compositional generalizability.

The overall framework of our proposed method is shown in Fig. 2. Given a video-query pair (V_p, Q_p), we first construct a set of hierarchical hard negative queries $\{Q_{hn}^i\}_{i=1}^3$ via a progressive mask-and-predict strategy (Sec. 3.2). Notably, we utilize a large language model, namely GPT-3.5-Turbo [1], to select appropriate words from the training set for primitive replacement. This operation gradually changes the semantics of the original query while effectively avoiding implausible combinations. These manipulated queries, along with the original query Q_p and a negative query Q_n selected from other queries in the same mini-batch, constitute a set of queries. During the training phase, each video and its associated query set are first fed into a video encoder $\mathcal{F}_v(\cdot)$ and a text encoder $\mathcal{F}_t(\cdot)$ to extract corresponding features, which are then processed by the encoder to predict the saliency scores $\{S_p, S_{hn}^1, S_{hn}^2, S_{hn}^3, S_n\}$. Subsequently, we propose modeling the video-level saliency prior using a coarse-grained saliency ranking loss $\mathcal{L}_{cr}$, which incorporates two constraints designed to enhance the discriminability between positive and negative queries (Sec. 3.3). Concurrently, we employ a fine-grained saliency ranking loss $\mathcal{L}_{fr}$ to discern the saliency scores derived from the query set, which facilitates the learning of multi-granularity semantics by exploring the nuance among hierarchical negative samples. By combining $\mathcal{L}_{base}$ with the proposed coarse-to-fine saliency ranking loss (Sec. 3.4), our method can be seamlessly integrated into existing DETR-based models [18, 29], greatly enhancing their potential for compositional generalization.

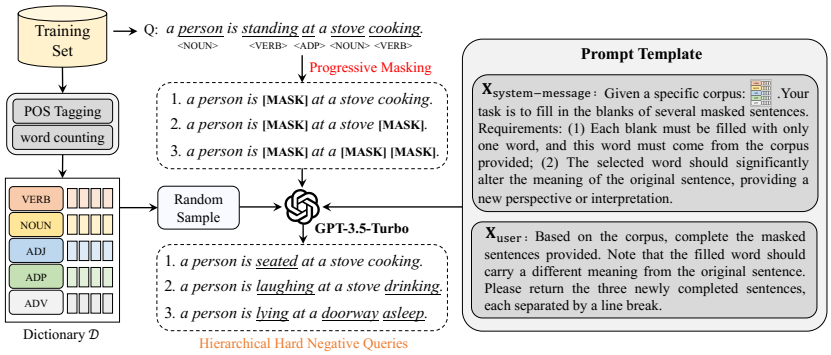


Fig. 3: The construction pipeline of hierarchical hard negative queries.

3.2 Hierarchical Hard Negatives Construction

Unlike previous methods that only consider verbs and nouns for negative query construction, we argue that other primitives like adverbs and prepositions also play an important role in composition generalization. Moreover, negative samples generated by randomly replacing primitives contain a large number of semantically infeasible combinations. Accordingly, we first propose an LLM-driven method to construct semantically viable hierarchical negative queries.

Fig. 3 shows the pipeline for constructing hierarchical negative queries. Specifically, we first use spaCy[3] to perform part-of-speech (POS) tagging on all query sentences in the training set, and construct a dictionary D by counting words from five types of primitives (verbs, nouns, adjectives, prepositions, and adverbs). Subsequently, for a given query Q, we progressively mask the primitives with different ratios in the original query according to their relative importance in linguistics, *i.e.*, verb-noun-adjective-preposition-adverb. Also, to ensure contextual consistency in negative queries, subjects (usually nouns) are only considered when other primitives are insufficient. Instead of filling in the masks with random selections from D, we resort to a powerful LLM, *i.e.*, GPT-3.5-Turbo, to generate semantically plausible hard negative queries. Considering the token limit inherent to LLMs, we randomly select a subset from the dictionary to ensure a balance between context and diversity for each sample. Furthermore, we have carefully crafted a prompt template to steer the LLM toward producing challenging negative queries. These negatives are semantically viable yet distinct from the original query, laying the foundation for fine-grained saliency ranking.

3.3 Coarse-to-Fine Saliency Ranking

The saliency scores measure the relevance of video clips to a given query, revealing the corresponding highlight moments. However, we observed that existing methods [18, 29] exhibit irrational saliency responses when faced with different

[3] spaCy: https://spacy.io/

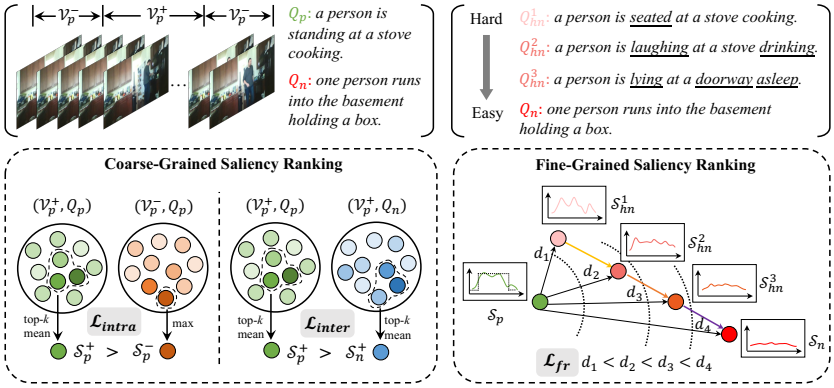


Fig. 4: An illustration of the Coarse-to-Fine Saliency Ranking strategy.

negative queries, as shown in Fig. 1. The saliency scores of some hard negative queries even surpass that of the original ones. This indicates that current methods fail to accurately match visual representations with corresponding primitives, struggling to discern the nuances between different queries. To this end, we introduce a coarse-to-fine saliency ranking strategy by establishing a multi-granularity semantic relationship between video clips and different negative queries. This approach enables the model to capture hierarchical semantic differences and enhances its compositional generalizability.

Coarse-Grained Saliency Ranking. As shown in Fig. 4, for a given video-text pair (V_p, Q_p), we designate video clips within and outside the ground-truth interval as V_p^+ and V_p^-, respectively, and Q_p as the corresponding positive query. Concurrently, a negative query is randomly selected from the same mini-batch, denoted as Q_n. Intuitively, the saliency scores within the ground-truth interval should surpass those outside, and scores elicited by positive queries should exceed those by negative ones. With these two priors, we introduce a coarse-grained ranking loss with dual constraints, formulated as:

$$\mathcal{L}_{cr} = \underbrace{\max(0, h_1 + S_p^- - S_p^+)}_{\mathcal{L}_{intra}} + \underbrace{\max(0, h_2 + S_n^+ - S_p^+)}_{\mathcal{L}_{inter}}, \tag{1}$$

$$S^+ = \frac{1}{k}\sum_{i=1}^{k} \text{sort}(S)_{T^+}, k = \max(1, \lfloor T^+/q \rfloor), \tag{2}$$

where S_p^+ and S_n^+ represent the top-k mean value of saliency scores yielded by (V_p^+, Q_p) and (V_p^+, Q_n), respectively, and S_p^- is the maximum of saliency scores produced by (V_p^-, Q_p). T^+ denotes the number of clips within the ground-truth interval, and q is a factor to control the selection ratio. h_1 and h_2 are two predefined margins. Notably, our constraints differ from [18] in two key ways: (1) Rather than using a single maximum value, we adapt to intervals of different scales based on interval length and q. (2) Beyond considering the

internal difference of positive queries, we also enlarge the saliency gap between positive and negative queries to achieve better discriminability.

Fine-Grained Saliency Ranking. While the coarse-grained ranking loss improves the discriminative capability of the video-text representation, it does not fully capture the relationships between the query primitives and the video clip. In DETR-based architectures, saliency scores are temporally aligned with the timestamps of localization boundaries, mirroring the relevance of the current video clip to the query. We argue that saliency scores tied to the original query should be temporally consistent with the ground truth, whereas those related to hard negatives ought to display a hierarchical disparity from the positive one, as illustrated in Fig. 4. Based on this assumption, we further propose a fine-grained ranking loss to refine the saliency scores derived from varying semantic levels of negative queries, formulated as:

$$\begin{aligned}\mathcal{L}_{fr} = &\max(0, m_0 + d(Y, S_p) - d(S_p, S^1_{hn})) \\ &+ \max(0, m_1 + d(S_p, S^1_{hn}) - d(S_p, S^2_{hn})) \\ &+ \max(0, m_2 + d(S_p, S^2_{hn}) - d(S_p, S^3_{hn})) \\ &+ \max(0, m_3 + d(S_p, S^3_{hn}) - d(S_p, S_n)),\end{aligned} \tag{3}$$

$$d(y, \hat{y}) = -\frac{1}{T}\sum_{i=1}^{T} y_i \log(\hat{y}_i), \tag{4}$$

where $\{S^i_{hn}\}^3_{i=1}$ denotes saliency scores of hierarchical negative queries, $d(\cdot)$ symbolizes the negative log-likelihood between the observation y and the prediction $\hat{y}$, measuring the disparity in the distribution of saliency scores across the temporal dimension T. In particular, due to the lack of ground-truths for the saliency scores, a value of 1 is assigned to moments inside the localization interval, and 0 is assigned to moments outside the interval, with Y denoting the pseudo saliency score. m_0 to m_3 represent four predefined margins. This loss not only underscores the nuanced semantics between the video and the query sentence, but also the differences in temporal distribution and magnitude of the saliency scores in the hard negative samples. By hierarchically constraining the saliency scores of these hard negatives, our method helps the model discern the nuances between various primitive words and video moments, suppressing irrational saliency responses, and further improving its capability to identify novel combinations.

3.4 Model Training Objectives

Since our method can be seamlessly integrated into existing DETR-based models, during the training phase, we optimize the model utilizing three distinct loss functions, with the overall objective expressed as:

$$\mathcal{L} = \mathcal{L}_{base} + \alpha\mathcal{L}_{cr} + \beta\mathcal{L}_{fr} \tag{5}$$

where $\mathcal{L}_{base}$ represents the basic loss of the DETR-based model, typically including bipartite matching loss, moment localization loss, and saliency loss. α

and β are two coefficients that control the loss weights. By combining the two constraints with the basic loss function, our method can significantly enhance the model's compositional generalization capabilities while preserving accuracy for in-distribution samples, as demonstrated in subsequent experiments.

4 Experiments

Datasets and Evaluation Metrics. We evaluate our method on two newly proposed benchmarks, Charades-CG and ActivityNet-CG [20], originated from Charades-STA [8] and ActivityNet Captions [17]. Each dataset is reorganized into four splits: Training/Test-Trivial/Novel-Composition/Novel-Word, where the latter three splits evaluate the model's performance on IID samples, novel combinations of seen words, and unseen words, respectively. In particular, the Novel-Composition split considers five types of new compositions: verb-noun, noun-noun, verb-adverb, adjective-noun, and preposition-noun. Following previous works [19, 20, 42], we use two main metrics to evaluate our methods, *i.e.*, "R@n, IoU = m" and mean Intersection over Union (mIoU).

Implementation Details. We adopt Moment-DETR [18] and QD-DETR [29] as our baselines and integrate them with our method using their officially released code. Unless specifically noted, other hyperparameters follow their default settings. Following [19, 42], for the Charades-CG dataset, we use pretrained I3D Network [3] to extract 1024-dimensional features for each 0.5-second clip. For ActivityNet-CG, we employ pretrained C3D network [16] to extract 4096-dimensional features for 1-second video clips. For text features, we follow [18, 29] to extract CLIP [31] features with 512 dimensions for each query. In hierarchical hard negative construction, we progressively mask the original query in Charades-CG at ratios of 25%, 50%, and 75%, while the masking ratios for ActivityNet-CG are set to 10%, 30%, and 50%. For Charades-CG, we set the learning rates for QD-DETR and Moment-DETR to 0.0001 and 0.0002, respectively, while for ActivityNet-CG, the learning rates for both models are set to 0.0002. The coarse-grained margins h_1 and h_2 are set to 1.0 and 2.0, respectively, while the relative thresholds m_0 to m_3 are set to 0.25. The factor q in Eq. (2) are set to 8 for both Charades-CG and ActivityNet-CG. All experiments are run on a single NVIDIA A100 GPU with a batch size of 32 training for 200 epochs.

4.1 Comparisons with the State-of-the-arts

Tab. 1 shows the overall performance of our approach on the Charades-CG dataset. We observe that: (1) While the baseline QD-DETR [29] outperforms the latest state-of-the-art methods (*e.g.*, SSL [19] and DeCo [42]) in the Test-Trivial spilt, there is still a performance gap in the Novel-Composition and Novel-Word splits, indicating that it has poor compositional generalization capabilities. (2) Our method can significantly improve the compositional generalizability of QD-DETR, elevating 7.93% and 6.60% in R1@0.5 and R1@0.7 in Novel-Composition split, respectively, finally outperforming on all three test splits. (3) Our method

Table 1: Performance (%) of state-of-the-art methods on the Charades-CG dataset. The best result is shown in **bold** and the second best is underlined. 'WS': weakly-supervised methods. 'RL': reinforcement learning methods. 'PB': proposal-based methods. 'PF': proposal-free methods. † indicates the results of our implementation using the officially released code. ⋆ denotes the results relying on external detector knowledge.

Setting	Method	Test-Trivial			Novel-Composition			Novel-Word		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
WS	WSSL [6]	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL	TSP-PRL [36]	39.86	21.07	38.41	16.3	2.04	13.52	14.83	2.61	14.03
PB	TMN [23]	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN [52]	48.06	27.10	43.72	32.74	15.25	31.5	37.12	18.99	35.04
	2D-TAN+SSL [19]	53.91	31.82	46.84	35.42	17.95	33.07	43.60	25.32	39.32
	MS-2D-TAN [51]	57.85	37.63	50.51	43.17	23.27	38.06	45.76	27.19	40.80
	MS-2D-TAN+SSL [19]	58.14	37.98	50.58	46.54	25.10	40.00	50.36	28.78	43.15
PF	LGI [30]	49.45	23.8	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VLSNet [50]	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	VISA⋆ [20]	53.20	26.52	47.11	45.41	22.71	42.03	42.35	20.88	40.18
	Deco [42]	58.75	28.71	49.06	47.39	21.06	40.70	-	-	-
	Moment-DETR† [18]	49.48	28.04	44.82	39.42	18.62	36.61	46.76	24.75	41.70
	Moment-DETR+Ours	57.14	33.85	49.32	44.65	23.21	39.86	47.05	24.32	41.57
	QD-DETR† [29]	59.24	33.43	50.92	42.30	21.09	38.55	46.04	26.33	42.89
	QD-DETR+Ours	**60.66**	**38.60**	**52.53**	**50.23**	**27.69**	**44.14**	**55.25**	**35.25**	**48.10**

can be integrated into existing DETR-based models to unlock their compositional generalizability. In Tab. 2, we also achieve competitive results on the ActivityNet-CG dataset across two baselines. In the Novel-Composition split, our method promotes the performance of QD-DETR by 2.65%, 3.43%, and 1.43% in R1@0.5, R1@0.7 and mIoU, respectively. Additionally, by integrating our approach, Moment-DETR's performance experiences a boost of 1.39%, 0.69%, and 1.50% in R1@0.5, R1@0.7 and mIoU metrics, respectively. Notably, VISA [20] leverages external knowledge from off-the-shelf object detectors and action recognition models while our method is conducted in an end-to-end manner. Although our method underperforms VISA [20] in the Test-Trivial split, it still achieves comparable performance in the Novel-Composition split, which proves that our method has a better capability of compositional generalization.

4.2 Ablation Studies

We further provide ablation studies to validate the effectiveness of the proposed method, including various constraints in the coarse-to-fine saliency ranking loss, diverse hierarchical negative queries, and several hyperparameter settings. We use QD-DETR as the baseline to explore the insights.

Coarse-Grained Saliency Ranking. We report the contributions of each constraint within the coarse-grained saliency ranking loss in Tab. 3. Note that the proposed intra-ranking loss shares the same objective with the saliency loss in $\mathcal{L}_{base}$. Therefore, by replacing it with $\mathcal{L}_{intra}$, our loss yields improvements across

Table 2: Performance (%) of state-of-the-art methods on the ActivityNet-CG dataset. The best result is shown in **bold** and the second best is underlined. 'WS': weakly-supervised methods. 'RL': reinforcement learning methods. 'PB': proposal-based methods. 'PF': proposal-free methods. † indicates the results of our implementation using the officially released code. ⋆ denotes the results relying on external detector knowledge.

Setting	Method	Test-Trivial			Novel-Composition			Novel-Word		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
WS	WSSL [6]	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL	TSP-PRL [36]	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
PB	TMN [23]	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN [52]	**44.50**	**26.03**	42.12	22.80	9.95	28.49	23.86	10.37	28.88
PF	LGI [30]	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet [50]	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA⋆ [20]	47.13	29.64	44.02	31.51	16.73	35.85	30.14	15.90	35.13
	Deco [42]	43.98	24.25	43.47	27.35	11.66	31.27	-	-	-
	Moment-DETR† [18]	42.73	25.31	42.19	29.29	13.71	31.63	26.84	13.34	29.95
	Moment-DETR+Ours	44.19	25.81	**43.49**	**30.60**	**14.40**	**33.13**	**29.59**	**15.10**	**32.43**
	QD-DETR† [29]	41.80	20.88	41.15	26.91	10.96	31.01	27.09	11.38	31.21
	QD-DETR+Ours	43.76	25.98	42.86	29.56	14.37	32.44	27.60	13.11	30.98

Table 3: Ablation studies for coarse-grained ranking on Charades-CG dataset. * means that we replace the saliency loss in $\mathcal{L}_{base}$ with our $\mathcal{L}_{intra}$ and yield better performance.

$\mathcal{L}_{base}$	$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	Test-Trivial			Novel-Composition		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
✓			59.24	33.43	50.92	42.30	21.09	38.55
✓	✓*		60.24	35.89	51.73	44.02	22.84	39.23
✓		✓	60.17	37.11	51.96	**46.69**	24.87	41.74
✓	✓*	✓	**61.98**	**37.56**	**53.38**	46.25	**24.93**	**41.88**

all metrics on two test sets, with R1@0.7 increasing by 2.46% on Test-trivial and 1.75% on Novel-Composition. We also notice that the boosting effect of $\mathcal{L}_{inter}$ is more significant compared to $\mathcal{L}_{intra}$. By combining both with $\mathcal{L}_{base}$, the overall performance can be further increased, resulting in absolute gains of 2.46% and 3.33% for mIoU on Test-Trivial and Novel-Composition, respectively. The results show the $\mathcal{L}_{cr}$ enhances the discriminability between video clips inside and outside the ground truth interval, as well as between positive and negative query responses, simultaneously improving the compositional generalizability.

Fine-Grained Saliency Ranking. We present the effects of each constraint within the fine-grained saliency ranking loss in Tab. 4, and its synergy with coarse-grained ranking loss. We can observe that: (1) As fine-grained ranking constraints are gradually added, there is a general trend of improvement across all three metrics, among which $\mathcal{L}_{fr}^{1}$ plays a leading role, significantly improving R1@0.5 by 4.13%. (2) Without complete constraints, the introduction of $\mathcal{L}_{fr}^{3}$ leads to a slight performance degradation. This trend remains consistent both before and after combining with $\mathcal{L}_{cr}$. Interestingly, when all constraints

Table 4: Ablation studies for fine-grained saliency ranking constraints in the Novel-Composition split of Charades-CG. $\mathcal{L}_{fr}^{1}$ to $\mathcal{L}_{fr}^{4}$ are four constraints in order in $\mathcal{L}_{fr}$.

$\mathcal{L}_{base}$	$\mathcal{L}_{fr}^{1}$	$\mathcal{L}_{fr}^{2}$	$\mathcal{L}_{fr}^{3}$	$\mathcal{L}_{fr}^{4}$	$\mathcal{L}_{cr}$	R1@0.5	R1@0.7	mIoU
✓						42.30	21.09	38.55
✓	✓					46.43	23.39	41.65
✓	✓	✓				46.72	24.35	42.29
✓	✓	✓	✓			46.40	23.30	41.88
✓	✓	✓	✓	✓		46.98	24.00	41.94
✓	✓	✓			✓	48.40	24.55	43.05
✓	✓	✓	✓		✓	47.41	25.68	41.52
✓	✓	✓		✓	✓	49.51	24.52	43.27
✓	✓	✓	✓	✓	✓	**50.23**	**27.69**	**44.14**

Table 5: Comparison of different large language models for hard negative construction in the Novel-Composition test split.

Dataset	Hard Negatives	R1@0.5	R1@0.7	mIoU
Charades-CG	random sample	47.41	25.33	42.50
	Llama 3 [28]	48.75	25.22	42.89
	Gemini-1.5 Flash [9]	48.69	25.60	43.54
	GPT-3.5 Turbo [1]	**50.23**	**27.69**	**44.14**
ActivityNet-CG	random sample	29.59	12.89	32.06
	Llama 3 [28]	29.24	13.56	31.82
	Gemini-1.5 Flash [9]	**29.72**	13.24	31.96
	GPT-3.5 Turbo [1]	29.56	**14.37**	**32.44**

Table 6: Contribution of including *prepositions* and *adverbs* into hard negatives on Charades-CG Novel Composition split.

Method	R1@0.5	R1@0.7	mIoU
MD+Ours w/o *prep* & *adv*	43.03	22.08	38.79
MD+Ours	44.65	23.21	39.86
QD+Ours w/o *prep* & *adv*	48.87	25.28	43.30
QD+Ours	**50.23**	**27.69**	**44.14**

Table 7: Performance (mIoU) of our method on different composition types of Charades-CG Novel Composition split.

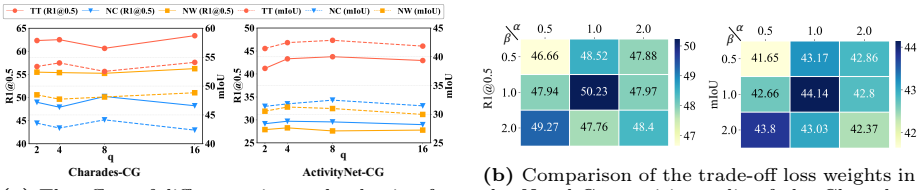
Method	verb-noun	adj-noun	noun-noun	verb-adv	prep-noun
MD	36.01	28.79	41.95	34.45	35.38
MD+Ours	40.29	37.54	41.56	37.69	38.48
QD	37.51	41.14	44.70	34.65	33.72
QD+Ours	43.25	45.34	45.53	41.05	38.71

are integrated into the baseline, $\mathcal{L}_{fr}^{3}$ further improves the performance by 3.17% in R1@0.7, which suggests that it only realizes full potential when forming a complete hierarchy of constraints.

Comparison of Hierarchical Negative Queries. We compare the effects of the random sample-based and LLM-based hard negative queries in Tab. 5. From the Charades-CG results, we observe that LLM-based negative queries significantly outperform random sampling, with improvements of 2.82% on R1@0.5 and 1.64% on mIoU. Among all the evaluated LLMs, the negative samples generated by GPT-3.5 Turbo perform better. We also note that the improvement using the LLM-based approach over random sampling is less significant in ActivityNet-CG. A possible reason is that the query sentences are longer while the replacement ratio is lower, so the retained context is still sufficient for localization. However, the LLM-based approach consistently outperforms random sampling in R1@0.7, indicating its potential for precise grounding.

Ablation of Different Composition Types. We conduct an ablation study by excluding *prepositions* and *adverbs* for hard negative construction. Tab. 6 shows that considering *prepositions* and *adverbs* effectively improves the model's perception of non-dominant primitives. Tab. 7 further shows that our method consistently improves the compositional generalizability of existing DETR-based methods across different composition types, validating the rationality of the proposed hard negative construction method.

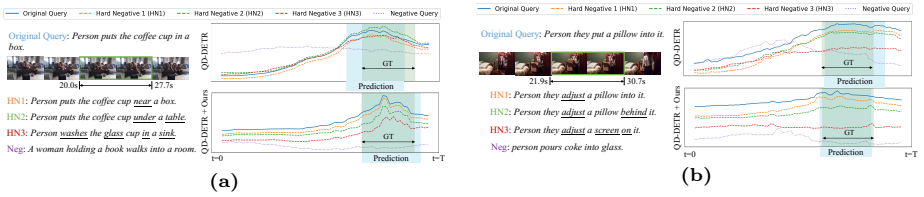
Hyperparameter Evaluation. In Fig. 5a, we explore the effect of different q on the top-k selection in Eq. (2). For Charades-CG, with the increase of q,



(a) The effect of different q in top-k selection for coarse-grained ranking loss.

(b) Comparison of the trade-off loss weights in the Novel-Composition split of the Charades-CG Dataset.

Fig. 5: Hyperparameter Evaluation.



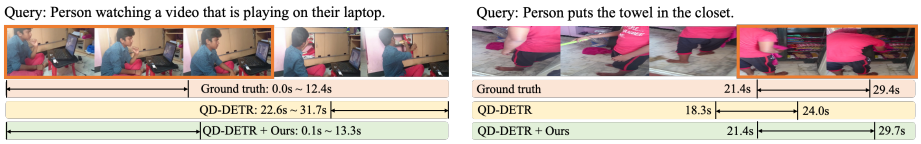
(a) **(b)**

Fig. 6: Visualization of saliency scores given different query sentences. (a) and (b) are test samples from Charades-CG.

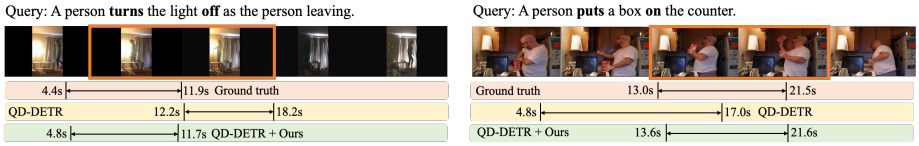
the metrics including R1@0.5 and mIoU on Test-Trivial and Novel-Composition show an opposite trend, indicating a competitive balance between the two splits. For ActivityNet-CG, the trends for Test-Trivial and Novel-Composition are consistent, while larger q leads to a decrease in performance on all three splits. When $q = 8$, the model achieves the best Novel-Composition results on both datasets. The heatmaps in Fig. 5b illustrate the effects of different loss weights α and β in Eq. (5) in the Novel-Composition split. We empirically find that the optimal performance in R1@0.5 and mIoU is achieved when both α and β are set to 1.0. Therefore, we use this setting by default in our experiments.

4.3 Qualitative Analysis

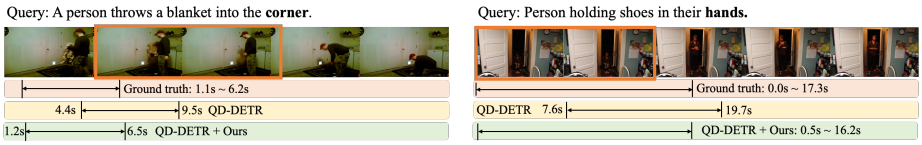
We visualize the saliency scores of several cases in Charades-CG in Fig. 6 and observe that the existing work struggles with hard negative queries, showing irrational saliency responses. For instance, in Fig. 6a, the hard negative query "*Person puts the coffee cup under a table.*" is even more salient than the positive query "*Person puts the coffee cup in a box*", leading to imprecise moment localization. In contrast, our approach consistently improves the model's ability to distinguish different words between positive and hard negative queries and yield hierarchical responses, thereby achieving better moment localization and compositional generalization. Additionally, our method can also catch the nuanced variation of adverbs and prepositions, such as different prepositions "into" and "behind" in Fig. 6b. This indicates that ours are more sensitive to the semantic changes of different non-dominant primitives.



(a) Two Samples from the Test-Trivial split



(b) Two Samples from the Novel-Composition split



(c) Two Samples from the Novel-Word split

Fig. 7: Qualitative comparison in different test splits of Charades-CG.

Fig. 7 illustrates several qualitative examples in three splits of Charades-CG. In the Test-Trivial split, although the queries don't contain unseen compositions and words, our method demonstrates more precise alignment than the baseline. When encountering the Novel-composition "**turns** the light **off**" and "**put** a box **on**" and Novel-Word "**corner**" and "**hands**" in the queries, our method can still generalise well to them. The presented results indicate that our method effectively guides DETR-based models in utilizing hierarchical negative samples to enhance the generalizability of unseen compositions and unseen words.

5 Conclusion

In this paper, we propose SHINE, a Saliency-aware HIerarchical NEgative ranking method for compositional temporal grounding. We first utilize an LLM to produce semantically plausible yet distinct hierarchical hard negatives from the original query. Furthermore, we introduce a coarse-to-fine saliency ranking strategy that establishes a multi-granularity semantic relationship between video and hard negatives. Extensive experiments demonstrate that SHINE substantially enhances the compositional generalization capabilities of current DETR-based temporal grounding models.

Acknowledgements. This work was partially supported by Veritone and Adobe and utilised Queen Mary's Apocrita HPC facility from QMUL Research-IT. This work was also partially supported by the Office of Naval Research (ONR) grant (N00014-23-1-2417 & N00014-23-1-2046), National Science Foundation (NSF) CAREER under award 2028626, and NSF SaTC Award 1949694. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 5, 12
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 9
4. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 162–171 (2018) 4
5. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., et al.: Dense and aligned captions (dac) promote compositional reasoning in vl models. Advances in Neural Information Processing Systems **36** (2023) 2, 4
6. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. Advances in Neural Information Processing Systems **31** (2018) 10, 11
7. Gandhi, M., Gul, M.O., Prakash, E., Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Measuring compositional consistency for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5046–5055 (2022) 4
8. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017) 1, 3, 4, 9
9. Google: Gemini-1.5 flash (2024), https://deepmind.google/technologies/gemini/flash/ 12
10. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11287–11297 (2021) 4
11. Hao, J., Sun, H., Ren, P., Wang, J., Qi, Q., Liao, J.: Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In: European Conference on Computer Vision. pp. 130–147. Springer (2022) 4
12. Hou, Z., Zhong, W., Ji, L., Gao, D., Yan, K., Chan, W.K., Ngo, C.W., Shou, Z., Duan, N.: Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. arXiv preprint arXiv:2209.10918 (2022) 4
13. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in Neural Information Processing Systems **36** (2023) 4
14. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13846–13856 (2023) 3, 4
15. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 4

16. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014) 9
17. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017) 1, 3, 9
18. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021) 1, 3, 4, 5, 6, 7, 9, 10, 11
19. Li, C., Li, Z., Jing, C., Jia, Y., Wu, Y.: Exploring the effect of primitives for compositional generalization in vision-and-language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19092–19101 (2023) 2, 4, 9, 10
20. Li, J., Xie, J., Qian, L., Zhu, L., Tang, S., Wu, F., Yang, Y., Zhuang, Y., Wang, X.E.: Compositional temporal grounding with structured variational cross-graph correspondence learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3032–3041 (2022) 2, 4, 9, 10, 11
21. Li, X., Yang, X., Wei, K., Deng, C., Yang, M.: Siamese contrastive embedding network for compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9326–9335 (2022) 4
22. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023) 4
23. Liu, B., Yeung, S., Chou, E., Huang, D.A., Fei-Fei, L., Niebles, J.C.: Temporal modular networks for retrieving complex compositional activities in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 552–568 (2018) 10, 11
24. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 843–851 (2018) 4
25. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3042–3051 (2022) 3, 4
26. Lu, X., Guo, S., Liu, Z., Guo, J.: Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23560–23569 (2023) 4
27. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023) 4
28. Meta: Llama 3 (2024), https://llama.meta.com/llama3/ 12
29. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023) 1, 2, 3, 4, 5, 6, 9, 10, 11
30. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020) 4, 10, 11

31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 9
32. Ray, A., Radenovic, F., Dubey, A., Plummer, B., Krishna, R., Saenko, K.: cola: A benchmark for compositional text-to-image retrieval. Advances in Neural Information Processing Systems **36** (2023) 4
33. Singh, H., Zhang, P., Wang, Q., Wang, M., Xiong, W., Du, J., Chen, Y.: Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 869–893 (2023) 2
34. Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., Soatto, S.: Linear spaces of meanings: compositional structures in vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15395–15404 (2023) 2, 4
35. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2613–2623 (2022) 4
36. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12386–12393 (2020) 10, 11
37. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2986–2994 (2021) 4
38. Xu, G., Chai, J., Kordjamshidi, P.: Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5774–5783 (2024) 4
39. Xu, G., Kordjamshidi, P., Chai, J.: Metarevision: Meta-learning with retrieval for visually grounded compositional concept acquisition. arXiv preprint arXiv:2311.01580 (2023) 4
40. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9062–9069 (2019) 4
41. Yan, S., Xiong, X., Nagrani, A., Arnab, A., Wang, Z., Ge, W., Ross, D., Schmid, C.: Unloc: A unified framework for video localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13623–13633 (2023) 4
42. Yang, L., Kong, Q., Yang, H.K., Kehl, W., Sato, Y., Kobori, N.: Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23130–23140 (2023) 2, 4, 9, 10, 11
43. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 982–990 (2016) 3
44. Yu, Z., Zheng, L., Zhao, Z., Wu, F., Fan, J., Ren, K., Yu, J.: Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23191–23200 (2023) 4

45. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Advances in Neural Information Processing Systems **32** (2019) 4
46. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9159–9166 (2019) 4
47. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022) 2, 4
48. Zeng, Y., Cao, D., Wei, X., Liu, M., Zhao, Z., Qin, Z.: Multi-modal relational graph for cross-modal video moment retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2215–2224 (2021) 4
49. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1247–1257 (2019) 4
50. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020) 1, 10, 11
51. Zhang, S., Peng, H., Fu, J., Lu, Y., Luo, J.: Multi-scale 2d temporal adjacency networks for moment localization with natural language. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9073–9087 (2021) 10
52. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020) 1, 4, 10, 11
53. Zheng, Z., Zhu, H., Nevatia, R.: Caila: Concept-aware intra-layer adapters for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1721–1731 (2024) 4
54. Zhu, Z., Tang, W., Wang, L., Zheng, N., Hua, G.: Enriching local and global contexts for temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13516–13525 (2021) 1