2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00263

Atomic-action-based Contrastive Network for Weakly Supervised Temporal Language Grounding

Hongzhou Wu†
Parallel and Distributed Processing Lab
National University of Defense Technology
Changsha, China
whz@nudt.edu.cn

Yifan Lyu†
Institute of Software
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
yifan2018@iscas.ac.cn

Xingyu Shen
Parallel and Distributed Processing Lab
National University of Defense Technology
Changsha, China
shenxingyu17@nudt.edu.cn

Xuechen Zhao
School of Computer
National University
of Defense Technology
Changsha, China
zhaoxuechen@nudt.edu.cn

Mengzhu Wang
Parallel and Distributed Processing Lab
National University of Defense Technology
Changsha, China
dreamkily@gmail.com

Xiang Zhang*
Institute for Quantum Information and the State Key Lab
of High Performance Computing
National University of Defense Technology
Changsha, China
zhangxiang08@nudt.edu.cn

Zhigang Luo
Parallel and Distributed Processing Lab
National University of Defense Technology
Changsha, China
zgluo@nudt.edu.cn

***Abstract*—As one knows, an event often consists of several actions while each action is atomic. Inspired by this insight, we propose a novel framework named Atomic-action-based Contrastive Network model (ACN) for weakly supervised temporal language grounding task to localize the query-related event moment in an untrimmed video, without access to any temporal annotations. Specifically, ACN first determines the accurate moment boundary of each action in a query-agnostic way. This can adequately exploit homogeneous visual cues while impeding the heterogeneity of the query from hurting the atomicity of visual action, *i.e.*, action boundary. To effectively localize the query-related event, we seek the discriminative words in the given query, and explore a composite-grained contrastive module to retrieve those corresponding atomic actions in the common latent space across modalities. This boosts feature discrimination of visual event segment to remove irrelevant action video segments. Experiments on two popular datasets show the efficacy of our model.**

***Index Terms*—weakly supervised temporal language grounding, cross-modal interaction, contrastive learning, atomic action, discriminative word**

I. Introduction

Given a natural language description and an untrimmed video, temporal language grounding (TLG) aims to predict the temporal boundary of the video event semantically corresponding to the given sentence. In this respect, many approaches like [1]–[4] focus on the full-supervised setting which annotates start time and end time of the matching video segment. However, it is expensive and time-consuming to annotate the temporal boundary of the matching video segment. Moreover, such annotations are ambiguous and noisy because of the subjectivity of annotaters. To this end, some researchers are dedicated to weakly supervised setting of TLG, which only uses the video-sentence pairs for training, without the help of any temporal annotations. Most existing weakly supervised approaches [5]–[9] treat this task as a multiple instance learning (MIL) problem. Recent methods [10], [11] extensively ensure the predicted moment to better reconstruct the query. However,the boundary of the predicted moment is still inexact because the affinity or similarity values between video frame features and the query feature seem less discriminative, as shown in Fig. 1 (a).

To address the above issue, we propose a novel Atomic-action-based Contrastive Network model (ACN) which applies the insight of video action recognition task for weakly supervised TLG. It is based on the general fact that each action is viewed to be atomic, while the atomicity of an action visually has the accurate moment boundary. That is, those consecutive video frames forming an action are intuitively indivisible and seem more distinguishable from those frames of other actions. Then it is relatively easier to identify the moment boundary of an action with visual cues than with the corresponding heterogeneous query semantics. Thus it could be preferable to pave the safe way for subsequent cross-modal

†These arthors are equally contributed to this work
*Xiang Zhang is the corresponding author

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00263

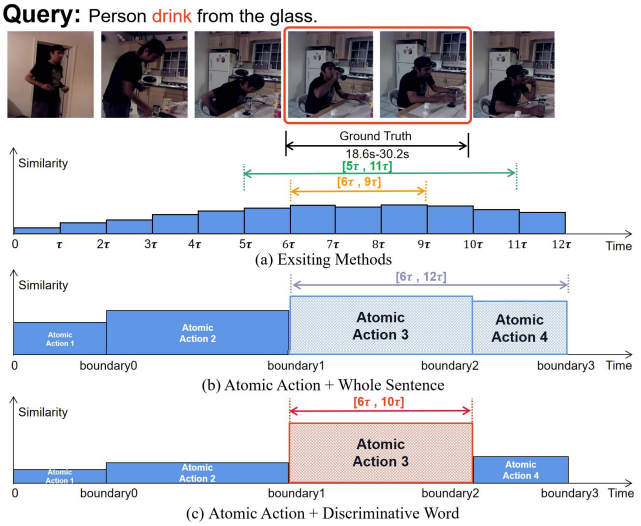

Fig. 1. (a) Most existing methods predict an inexact temporal boundary because in the cross-modal alignment those consecutive frames have similar affinity for the query, due to weak feature discrimination. (b) Atomic action can recognize the boundary but the global sentence representation may weaken the importance of some critical words like "drink". (c) The joint merits of atomic actions and discriminative word can remove many irrelevant actions.

semantic alignment by in advance recognizing and organizing such atomic actions in a query-agnostic way. In fact, modality-specific visual features originally trained for action recognition can be responsible for grouping the video frames into several atomic actions, and interestingly, so do the pre-trained CLIP [12] features. Hence, this way can fully adopt homogeneous visual cues while keeping the heterogeneity impact of the sentence like cross-modal semantic misalignment from hurting the atomicity of visual action. In other words, the atomic actions provide another reliable temporal supervision information, *i.e.*, accurate action boundary, to localize the event boundary, as Fig. 1 (b)-(c) shows. Moreover, this induces the global visual context of action representation. In (weakly) supervised TLG task, the query-related event is often a composition of some atomic actions, *e.g.,* drink from the glass, thus the event boundary localization equals to simultaneously localize several atomic actions. Note that Fig. 1 (b) has the inaccurate moment boundary of the event, because the event boundary not only relies on action boundary but also is closely relevant to the query.

Actually, the query used by previous studies primarily serves as the global semantic representation. This loses critical details of some key words like the *actions*. This is because those words indeed contain vital semantics of the descriptive event in the query but usually are weakened due to simple word representation average or concatenation. Currently several proposed approaches such as LoGAN [8] emphasize the word embedding and exploit the frame-word level matching for cross-modal alignment. Nonetheless, the affinity between a single frame and one word is easily misled without the help of the global context of visual cues. Worse still, this would directly destroy the visual atomicity of the action. To fully exploit query semantics and visual cues, we highlight discriminative words and leverage the global context of the detected atomic actions to perform the composite-grained constrative learning for feature discrimination. In detail, the coarse-grained contrastive learning matches the attentive visual features with the global sentence representation, while the fine-grained analogy aligns the detected atomic actions with discriminative words. In this way, our method enjoys the joint merits of the global and local context matching. Benefiting from both accurate action boundary and discriminative word, Fig. 1 (c) shows that our method can accurately localize the moment boundary of the query-related event.

In summary, the main contributions of our work are:

- We propose a novel Atomic-action-based Contrastive Network (ACN) for weakly supervised TLG, which considers the atomicity of actions to pave the safe way for refining the moment boundary of the event in a query-agnostic way.
- We devise the composite-grained contrastive module with a discriminative word strategy to enhance feature discrimination of atomic actions, thereby effective filtering out those video actions irrelevant to the query.
- Experiments on both Charades-STA and ActivityNet-Captions show that ACN achieves large performance gains as compared to recent well-established methods. This implies the potential of our insight in weakly supervised temporal language grounding.

II. Related Works

TLG seeks the video segment closely related to the nature language query. Most of previous works [1]–[4] are in fully supervised setting that relies on temporal annotations during training. Differently, weakly supervised TLG only has an access to video-sentence pairs available. Most approaches [5]–[9] in this respect are based on multi-instance learning (MIL) paradigm, which considers the mismatch videos within a batch as negative pairs to enhance cross-modal alignment. Another way is built on the reconstruction prior. For instance, SCN [11] first generates several proposals and then selects the top-k proposals which have the ability to reconstruct the masked words in the given query. Besides using the reconstruction mechanism, CNM [10] introduces the contrastive negative sample mining within the same video to further improve feature discrimination. However, the aforementioned methods are unaware of the boundary of the query-related event moment, because the similarity scores of consecutive video frames are often smooth without the help of any reliable temporal annotations during training. Thus, we will adopt the atomicity of actions as new priors to generate extra temporal supervision information to remedy this issue.

III. Approach

A. The Overall Architecture

Given an untrimmed video and a nature language sentence, a TLG model aims to localize the most relevant moment $\mathbf{T}$ =

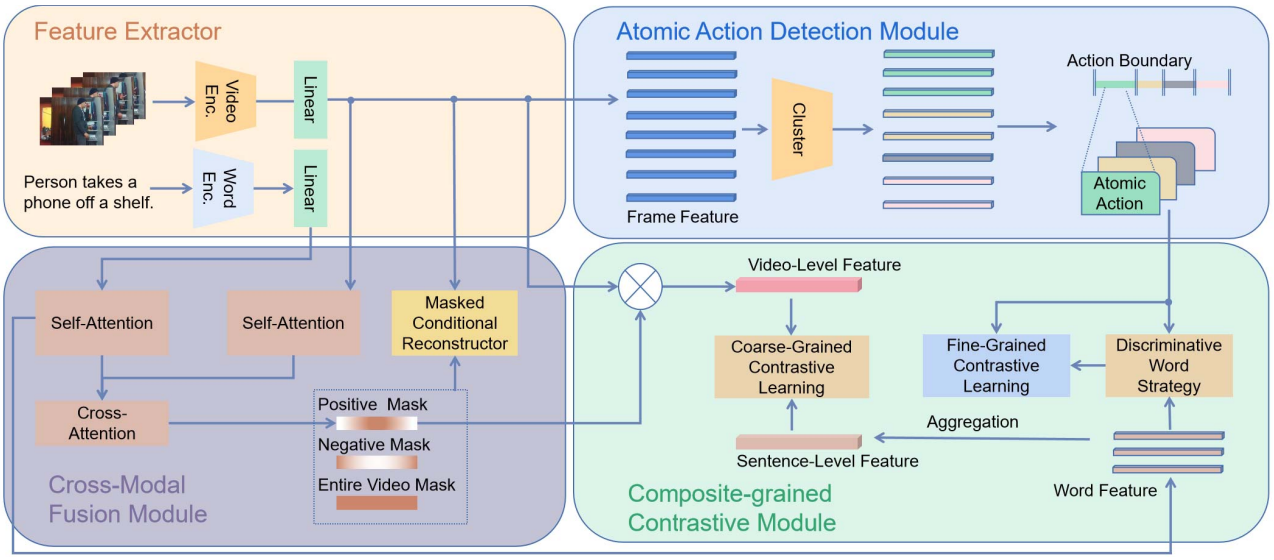

Fig. 2. The overall architechture of ACN. It consists of four main components: a feature extractor, a cross-modal fusion module, an atomic action detection module, and a composite-grained contrastive module.

(t_s, t_e).Here we consider the weakly supervised setting where the video-sentence pairs are only available. As shown in Fig. 2, the proposed model consists of four main components. The visual and word features are produced by a feature extractor, which will be fed to a cross-modal fusion module to generate the masks of positive and negative samples. We adopt a mask conditioned reconstructor using the attentive visual features with these generated masks for reconstructing the masked query to perform cross-modal interaction. Besides, an atomic action detection module directly groups the visual feature into several atomic actions. Built off such atomic actions, a composite-grained contrastive module with discriminative words is devised for cross-modal semantic alignment by enhancing feature discrimination.

B. *Feature Extractor*

For a given untrimmed video, we extract its visual features $V = \{v_i\}_{i=0}^{L_v} \in \mathbb{R}^{L_v \times D}$ by a pre-trained vision backbone, where D means the dimension of features and L_v means the number of video frames. For the input query, we use the GloVe word2vec model [13] for feature extraction. Both embeddings will be fed into a fully-connected layer, which outputs feature $S = \{s_i\}_{i=0}^{L_s} \in \mathbb{R}^{L_s \times D}$ wherein L_s is the word size.

C. *Cross-Modal Fusion Module*

We design a cross-modal fusion module to perform interaction between visual and textual features. A multi-head attention function maps a query and a key-value pair as an output:

$$Att(Q, K, V) = \sum_{i=0}^{n} softmax(\frac{Q_i {K_i}^T}{\sqrt{d_k}})V_i \tag{1}$$

where Q, K, V indicates the query, key and value vectors, respectively, and d_k means the dimension of such vectors. The uni-modal multi-head attention will be applied for each modality to capture the corresponding context:

$$S' = Att(\mathbf{W}^{Q_s}S, \mathbf{W}^{K_s}S, \mathbf{W}^{K_s}S) \tag{2}$$

$$V' = Att(\mathbf{W}^{Q_v}V, \mathbf{W}^{K_v}V, \mathbf{W}^{K_v}V) \tag{3}$$

where $\mathbf{W}^{Q_s}$, $\mathbf{W}^{K_s}$, $\mathbf{W}^{Q_v}$, and $\mathbf{W}^{K_v}$ are learnable parameters. To fuse visual and query features, we introduce a cross-modal multi-head attention as:

$$H = Att(V', S', S') \tag{4}$$

As H contains the information from two modalities, it is fed into a grounding head to predict the center c and the width w of target moment $\mathbf{T} = (t_s, t_e)$. The (t_s, t_e) can be calculated by:

$$t_s = max(0, c - \frac{w}{2})L_v, \ t_e = min(1, c + \frac{w}{2})L_v \tag{5}$$

Following [10], the Gaussian mask of the positive sample m^p is generated by:

$$m^p = exp(-\frac{\alpha(i/L_v - c)^2}{2w^2}), i = 0, 1, 2, 3, ...L_v \tag{6}$$

where α is a hyperparameter used for controlling the variance of Gaussian curve. To promote a better cross-modal interaction, we follow CNM [10] to use the mask conditioned reconstructor with the mask of the positive sample m^p, the mask of the entire video $m^h = [1, 1, 1, ...]$ and the mask of the negative sample $m^e = 1 - m^p$. Accordingly, the reconstruction loss and intra-video contrastive loss are formulated as:

$$\mathcal{L}_{rec} = \mathcal{L}_{ce}^{p} + \mathcal{L}_{ce}^{h}, \tag{7}$$

$$\mathcal{L}_{ivc} = \max(0, \mathcal{L}_{ce}^{p} - \mathcal{L}_{ce}^{h} + \beta_1) + \max(0, \mathcal{L}_{ce}^{p} - \mathcal{L}_{ce}^{e} + \beta_2) \tag{8}$$

where $\mathcal{L}_{ce}^{p}$, $\mathcal{L}_{ce}^{h}$, $\mathcal{L}_{ce}^{e}$ is the cross-entropy loss calculated by the reconstructed query and the real word distribution. β_1, β_2 are hyperparameters satisfying $\beta_1 < \beta_2$.

D. Atomic Action Detection Module

The query-related event chiefly consists of several actions. Visually, they are in the form of an indivisible set of video frames, *i.e., the atomicity of action. That is, such consecutive video frames forming an atomic action are intuitively indivisible and seem more distinguishable from the frames of the other actions. Due to intra-domain feature homogeneity, it is preferable to use visual cues to group video frames into several atomic actions. This grouping process is called atomic action detection. To employ this insight*, we propose an atomic action detection module (AAD) to adaptively detect atomic actions in a query-agnostic way. All the frame features are grouped into M clusters $\mathbf{C} = \{C_0, C_1, ..., C_i\} \in \mathbb{R}^{M \times D}$ by using the k-means algorithm [22]:

$$l_i = \underset{0 \le j \le M}{argmin} \sum_{i=0}^{L_v} \sum_{v_i \in C_j} ||v_i - u_j||^2 \tag{9}$$

where l_i is the cluster label of v_i.

Since the change of visual features across different frames usually evolves smoothly over time, the nearby frames are often highly similar. If an abrupt change in visual feature space happens, the nearby frames ought to belong to two different clusters. Thus, such visual changes acts as the action boundary set as below:

$$\mathcal{B} = \{t | t : l_t \neq l_{t-1}\} \tag{10}$$

An atomic action consists of those frames belonging to the same cluster and thus can be represented by the corresponding cluster center u_j. Then we introduce the nearest boundary distance loss $\mathcal{L}_{NBD}$ as:

$$\mathcal{L}_{NBD} = \underset{t \in \mathcal{B}}{min} ||t_s - t|| + \underset{t \in \mathcal{B}}{min} ||t_e - t|| \tag{11}$$

Since $\mathcal{L}_{NBD}$ encourages our model to localize the nearest action boundary based on homogeneous visual cues only, it offers us a reliable temporal supervision, which is never used in weakly supervised setting.

E. Discriminative Word Strategy

To remove irrelevant actions, we propose a discriminative word strategy (DWS) to highlight those discriminative words which distinguish the query-related actions from similar background actions. We construct the most discriminative k words $\mathbf{W}^d = \{s_{d_1}, s_{d_2}, ..., s_{d_k}\}$ by:

$$\mathbf{W}^d = \underset{d_i \in [0, L_s]}{argmin} \sum_{i=0}^{k} \sum_{j=0}^{M} SC(u_j, s_{d_i}) \tag{12}$$

where $SC(u, s)$ denotes the cosine similarity between the features of atomic action u and word feature s.

F. Composite-Grained Contrastive Module

To build up cross-modal alignment, we perform a composite-grained contrastive learning (CGC) on inter-video and intra-video samples. Firstly, the coarse-grained contrastive learning (CG) is conducted for video-sentence alignment on inter-video samples. The video-level feature $\bar{p}$ is obtained by weighted averaging V with m_p, when the last hidden state $\bar{q}$ of S' is regarded as the global sentence feature. The coarse-grained contrastive loss can be formulated as:

$$\mathcal{L}_{cg} = -\log \frac{exp(\bar{p} \cdot \bar{q}/\tau)}{\sum_{i=1}^{\mathcal{N}} exp(\bar{p} \cdot \bar{q}_i/\tau)} \tag{13}$$

where $\bar{q}_i$ is the query corresponding to the other video within a batch. $\mathcal{N}$ is the batch size and τ is a temperature hyperparameter. $\mathcal{L}_{cg}$ indicates that the visual features and word features are mapped into a common latent space which is the basis of fine-grained alignment and matching.

Secondly, we conduct fine-grained contrastive learning (FG) by using atomic actions and the query with discriminative words. The entire video can be regarded as two parts: the set of atomic actions inside the positive sample is $\mathbf{A}_p$ and the set of those outside is $\mathbf{A}_n$. We apply the fine-grained loss as:

$$\begin{aligned} \mathcal{L}_{fg} = &max(0, MS(\mathbf{A}_p) - MS(\mathbf{A}_n) + \beta_3) \\ &+ max(0, AS(\mathbf{A}_p) - AS(\mathbf{A}_n) + \beta_4) \end{aligned} \tag{14}$$

$$MS(\mathbf{A}) = \underset{u \in \mathbf{A}}{max} \sum_{s_d \in \mathbf{W}^d} \frac{SC(u, s_d)}{|\mathbf{A}|} \tag{15}$$

$$AS(\mathbf{A}) = \sum_{u \in \mathbf{A}} \sum_{s_d \in \mathbf{W}^d} \frac{SC(u, s_d)}{|\mathbf{A}|} \tag{16}$$

where β_3 and β_4 are hyperparameters which control the margins. Thus, our total loss consists of all the losses above:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{cg} + \lambda_1 \mathcal{L}_{fg} + \lambda_2 \mathcal{L}_{NBG} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{ivc} \tag{17}$$

where λ_0, λ_1, λ_2, λ_3 and λ_4 are five hyperparameters which balance impacts of different losses.

IV. Experiments

A. Experimental Settings

To test the effectiveness of our approach, we perform experiments on two benchmark datasets: **Charades-STA** [1] is a dataset of indoor daily activities. It contains 12,408 moment-sentence pairs in the training set and 3,720 pairs in the test set. We report our results on the test split. **ActivityNet-Captions** [14] contains about 20k untrimmed videos and over 70k moment-sentence pairs. We use the val_1 split for validation and val_2 for testing

Following [10], we adopt the evaluation metric 'IoU=m' as our evaluation metric which means the percentage of predicted moments having Intersection over Union (IoU) larger than the threshold m. We also choose $m = \{0.3, 0.5, 0.7\}$ for Charades-STA, and $m = \{0.1, 0.3, 0.5\}$ for ActivityNet-Captions.

TABLE I
PERFORMANCE COMPARISON ON ACTIVITYNET DATASET ($IoU@m \in \{0.1, 0.3, 0.5\}$). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST ARE UNDERLINED.

Method	IoU = 0.1	IoU = 0.3	IoU = 0.5
DCCP [21]	-	41.6	23.2
WS-DEC [15]	62.71	41.98	23.34
EC-SL [20]	68.48	44.29	24.26
MARN [16]	-	47.01	29.95
SCN [11]	71.48	47.23	29.22
RTBPN [17]	73.73	49.77	29.63
WSLLN [6]	75.4	42.8	22.7
LCNet [9]	78.58	48.49	26.33
WSTAN [7]	79.78	52.45	30.01
CRM [18]	**81.61**	55.26	32.19
CNM [10]	78.13	55.68	33.33
Ours	78.78	**57.66**	**34.18**

TABLE II
PERFORMANCE COMPARISON ON THE CHARADES-STA DATASET ($IoU@m \in \{0.3, 0.5, 0.7\}$). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST ARE UNDERLINED.

Method	IoU = 0.3	IoU = 0.5	IoU = 0.7
TGA [5]	32.14	19.94	8.84
SCN [11]	42.96	23.58	9.97
DCCP [21]	-	29.8	11.9
WSTAN [7]	43.39	29.35	12.28
LoGAN [8]	48.04	31.74	13.71
MARN [16]	48.55	31.94	14.81
CRM [18]	53.66	34.76	16.37
LCNet [9]	59.60	**39.19**	**18.87**
RTBPN [17]	60.04	32.36	13.24
CNM [10]	60.04	35.15	14.95
Ours	**62.95**	37.02	15.26

B. Implementation Details

We pre-extract I3D visual features [19] for Charades-STA and CLIP features [12] for ActivityNet-Captions. For the input query, we apply the pre-trained GloVe [13] model to extract word embeddings with 300 dimensions.

The number of downsampled video clips is set to 200 for all datasets, and the max query length is set to 20. The number of cluster centers in kmeans is set to 32 while the number of the discriminative words is 2.The loss weights λ_0, λ_1 and λ_2 are set to 0.1, while λ_3 and λ_4 are set to 1.0 for all the datasets. During training, the batch size is set to 128 for Charades-STA and 256 for ActivityNet-Captions. The Adam optimizer with an initial learning rate of 0.0004 is adopted. The sensitive analysis on the number of atomic actions and the number of the discriminative words are reported in Appendix A.

C. Comparisons to the State-Of-The-Art

Tables I and II illustrate the performance comparisons of previous state-of-the-art methods on ActivityNet-Captions and Charades-STA. According to the results, we achieve the best weakly-supervised performance on IoU=0.3 while LCNet [9] performs better than ours on IoU=0.5 and IoU=0.7. However, we outperform than LCNet for ActivityNet-Captions dataset, which has larger vocabulary size and more diverse scenes than Charades-STA. For ActivityNet-Captions dataset, it is slightly inferior to CRM [18] on IoU=0.1, because CRM introduces extra paragraph descriptions with temporal ordering for all the scenes. However, ACN surpasses most methods on both IoU=0.3 and IoU=0.5, which hints the efficacy of our method.

TABLE III
THE IMPORTANCE OF THE ATOMIC ACTION DETECTION MODULE AND THE DISCRIMINATIVE WORD STRATEGY FOR LOCALIZATION ON CHARADES-STA DATASET.

Method	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
Full Model	**62.95**	**37.02**	**15.26**	**39.57**
w/o. AAD	59.12	35.97	15.23	38.06
w/o. DWS	60.83	35.47	14.76	38.39
w/o. Both	54.50	30.72	12.41	34.58

TABLE IV
THE EFFECTIVENESS OF THE COMPOSITE-GRAINED CONTRASTIVE MODULE FOR MOMENT LOCALIZATION ON CHARADES-STA DATASET.

CG	FG	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
•	•	**62.95**	**37.02**	**15.26**	**39.57**
•	×	61.43	35.94	15.14	38.59
×	•	61.18	36.16	14.53	38.76
×	×	60.13	35.56	14.76	38.17

D. Ablation Study

This section conducts the ablation study to analyze the importance of each component in our model. We also introduce mean Intersection over Union (mIoU) to report the results.

Effectiveness of the atomic action detection module and the discriminative word strategy. To analyze the effectiveness of both components, we construct the model with no use of them as the baseline. Compared to the baseline, the performance of the full model increases from 34.58% to 39.57% on mIoU. The obvious improvement shows that AAD and DWS have a beneficial impact on performance. As shown in the last two rows in Table III, the model only using AAD also surpasses the baseline by 3.81%, which suggests that AAD has an independent contribution to accurate moment boundary of the event. Moreover, DWS improves the performance by 3.48% on mIoU, in light of the second row and last row in Table III. This shows that the discriminative words can reduce the misalignment with the similar yet irrelevant video segments. The visualization of the localization results of the compared models can be found in Appendix B.

Effectiveness of the the composite-grained contrastive module. We analyze the effect of the coarse-grained and fine-contrastive learning on localization performance. From Table IV, we can find that the full model outperforms all the ablation models. It suggests that both coarse-grained and fine-grained contrastive learning are effective and complementary, because the former induces a common latent space across modalities while the latter can capture more fine-grained semantic information.

E. Qualitative Results

Fig. 3 shows the grounding results of the proposed method and recent approaches [10], [11]. From the subfigures (a) and

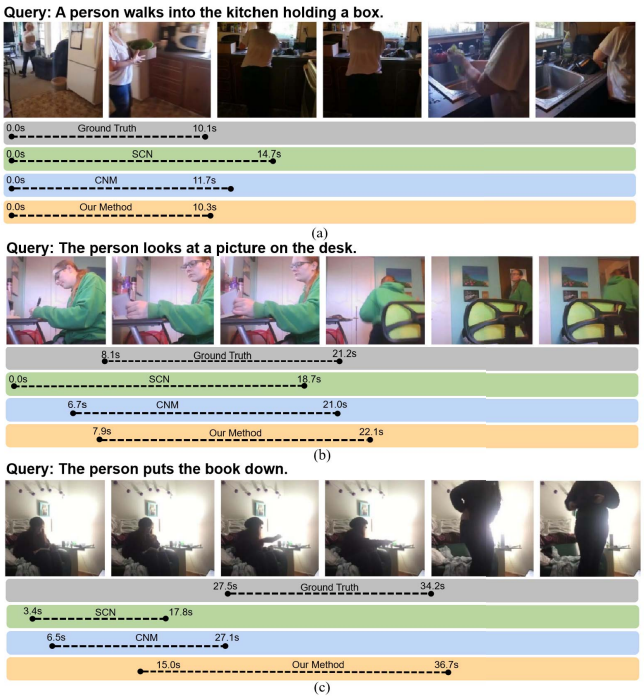

Fig. 3. The qualitative results of the proposed method.

(b), we predict the moment boundaries more precisely than previous methods, benefiting from temporal supervision of atomic action boundary. Moreover, Fig. 3 (c) illustrates that our method can distinguish video segments related to the query from the similar backgroundswhich mislead both SCN and CNM to make inexact predictions.

V. Conclusion

This paper proposes a novel weakly supervised TLG framework termed Atomic-action-based Contrastive Network (ACN). ACN can better determine the moment boundary of the event by the atomicity of actions, which provides alternative temporal supervision information for weakly supervised setting. Besides, the proposed composite-grained contrastive module with a discriminative word strategy is beneficial for removing redundant video actions similar to the background actions. Sound performance of our method on two popular datasets implies that our insight is feasible and promising.

References

[1] Gao, J., Sun, C., Yang, Z., Nevatia, R. (2017). "Tall: Temporal activity localization via language query". In Proceedings of the IEEE international conference on computer vision (pp. 5267-5275).

[2] Zhang, S., Peng, H., Fu, J., Luo, J. (2020, April). "Learning 2d temporal adjacent networks for moment localization with natural language". In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 12870-12877).

[3] Shen, X., Lan, L., Tan, H., Zhang, X., Ma, X., Luo, Z. (2022, June). "Joint Modality Synergy and Spatio-temporal Cue Purification for Moment Localization". In Proceedings of the 2022 International Conference on Multimedia Retrieval (pp. 369-379).

[4] Li, K., Guo, D., Wang, M. (2021, May). "Proposal-free video grounding with contextual pyramid network". In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 3, pp. 1902-1910).

[5] Mithun, N. C., Paul, S., Roy-Chowdhury, A. K. (2019). "Weakly supervised video moment retrieval from text queries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition" (pp. 11592-11601).

[6] Gao, M., Davis, L. S., Socher, R., Xiong, C. (2019). "Wslln: Weakly supervised natural language localization networks". arXiv preprint arXiv:1909.00239.

[7] Wang, Y., Deng, J., Zhou, W., Li, H. (2021). "Weakly supervised temporal adjacent network for language grounding". IEEE Transactions on Multimedia, 24, 3276-3286.

[8] Tan, R., Xu, H., Saenko, K., Plummer, B. A. (2021). "Logan: Latent graph co-attention network for weakly-supervised video moment retrieval". In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2083-2092).

[9] Yang, W., Zhang, T., Zhang, Y., Wu, F. (2021). "Local correspondence network for weakly supervised temporal sentence grounding". IEEE Transactions on Image Processing, 30, 3252-3262.

[10] Zheng, M., Huang, Y., Chen, Q., Liu, Y. (2022, June). "Weakly supervised video moment localization with contrastive negative sample mining". In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 3, pp. 3517-3525).

[11] Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H. (2020, April). "Weakly-supervised video moment retrieval via semantic completion network". In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11539-11546).

[12] Radford, Alec, et al. (2021, July). "Learning transferable visual models from natural language supervision". In International conference on machine learning (pp. 8748-8763). PMLR.

[13] Pennington, J., Socher, R., Manning, C. D. (2014, October). "Glove: Global vectors for word representation". In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[14] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, (2017). "Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision" (pp. 706-715)

[15] X. Duan, W. Huang, C. Gan, J. Wang, and J. Huang, "Weakly supervised dense event captioning in videos," Advances in Neural Information Processing Systems 31 (2018).

[16] Y. Song, J. Wang, L. Ma, Z. Yu, and J. Yu, (2020)"Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos," arXiv preprint arXiv:2003.07048.

[17] Z. Zhang, Z. Lin, Z. Zhao, J. Zhu, and X. He,(2020)"Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos", In Proceedings of the 28th ACM International Conference on Multimedia (pp. 4098-4106).

[18] J. Huang, Y. Liu, S. Gong, and H. Jin, (2021) "Cross-sentence temporal and semantic relations in video activity localization," In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7199-7208).

[19] J. Carreira and A. Zisserman, (2017) "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).

[20] Chen, S.; and Jiang, Y.-G. 2021. "Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8425–8435.

[21] Ma, F., Zhu, L. , Yang, Y. "Weakly Supervised Moment Localization with Decoupled Consistent Concept Prediction". Int J Comput Vis 130, 1244–1258 (2022).

[22] Zhao, Y. , Ming, Y. , Liu, X. , Zhu, E. , Yin, J. . (2018). ""Large-scale k-means clustering via variance reduction." Neurocomputing, 307.