Weakly-Supervised Temporal Article Grounding

Long Chen[♡], **Yulei Niu**[♡], **Brian Chen**[♡], **Xudong Lin**[♡], **Guangxing Han**[♡], **Christopher Thomas**[♢], **Hammad Ayyubi**[♡], **Heng Ji**[♠], **and Shih-Fu Chang**[♡]
[♡]Columbia University [♢]Virginia Tech [♠]University of Illinois at Urbana-Champaign
{cl3695, yn2338, bc2754, xl2798, gh2561, ha2578, sc250}@columbia.edu
chris@cs.vt.edu, hengji@illinois.edu

Abstract

Given a long untrimmed video and natural language queries, video grounding (VG) aims to temporally localize the semantically-aligned video segments. Almost all existing VG work holds two simple but unrealistic assumptions: 1) *All query sentences can be grounded in the corresponding video*. 2) *All query sentences for the same video are always at the same semantic scale*. Unfortunately, both assumptions make today's VG models fail to work in practice. For example, in real-world multimodal assets (*e.g.*, news articles), most of the sentences in the article can not be grounded in their affiliated videos, and they typically have rich hierarchical relations (*i.e.*, at different semantic scales). To this end, we propose a new challenging grounding task: *Weakly-Supervised temporal Article Grounding* (**WSAG**). Specifically, given an article and a relevant video, WSAG aims to localize all "groundable" sentences to the video, and these sentences are possibly at different semantic scales. Accordingly, we collect the first WSAG dataset to facilitate this task: **YouwikiHow**, which borrows the inherent multi-scale descriptions in wikiHow articles and plentiful YouTube videos. In addition, we propose a simple but effective method **DualMIL** for WSAG, which consists of a two-level MIL[1] loss and a single-/cross- sentence constraint loss. These training objectives are carefully designed for these relaxed assumptions. Extensive ablations have verified the effectiveness of DualMIL[2].

1 Introduction

Video Grounding (VG), *i.e.*, localizing video segments that semantically correspond to (coreference relation) query sentences, is one of the fundamental tasks in multimodal understanding. Further, video grounding can serve as an indispensable technique for many downstream applications, such as the text-oriented highlight detection (Lei et al., 2021), video retrieval (Miech et al., 2020) or video question answering (Ye et al., 2017; Xiao et al., 2022).

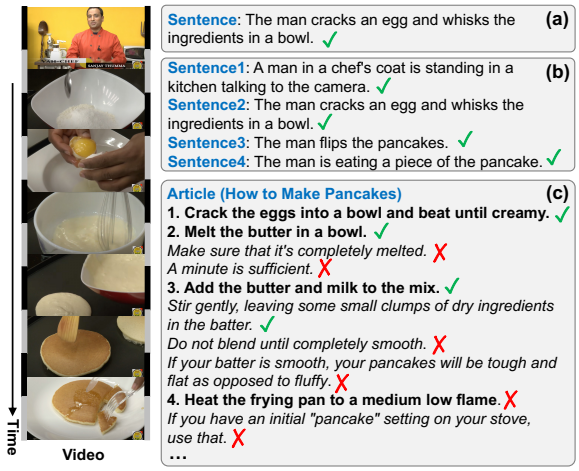

Figure 1: (a) Single sentence grounding: The query is a single sentence. (b) Multi-sentence grounding: The queries are multiple sentences. (c) Article grounding: The query is an article, which consists of multiple sentences at different scales (*e.g.*, *How to Make Pancakes*). **High-level** and *low-level* sentences are denoted with corresponding formats. ✓ and ✗ denote that sentence can or cannot be grounded to the video, respectively.

Early VG efforts mainly focus on single sentence grounding (Gao et al., 2017; Hendricks et al., 2017) (cf. Figure 1(a)). Thanks to advanced representation learning and multimodal fusion techniques, single sentence VG has achieved unprecedented progress over the recent years (Cao et al., 2021). The next step towards general VG is to ground multiple sentences to the same video (cf. Figure 1(b)). A straightforward solution for multi-sentence VG is utilizing the single sentence VG model for each sentence individually. Since these query sentences associated with the same video are always semantically related, recent multi-sentence VG methods directly ground all queries simultaneously by considering their temporal order or semantic relations (Bao et al., 2021; Shi et al., 2021).

Unfortunately, all existing VG attempts hold two

[1]MIL: Multiple Instance Learning.
[2]Codes: https://github.com/zjuchenlong/WSAG.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9402–9413
December 7-11, 2022

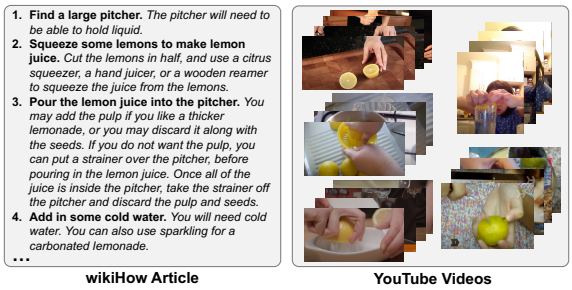

Figure 2: The only supervision for WSAG is a wikiHow article (*e.g.*, *How to Make Lemonade*) and some corresponding YouTube videos about the same task.

simple but unrealistic assumptions: 1) ***All query sentences can be grounded in the corresponding video***. Although this assumption is acceptable for the VG task itself, it greatly limits the usage of VG models in real-world multimodal assets. For example in news articles, most of the sentences in an article cannot be grounded in their affiliated videos. 2) ***All query sentences for the same video are always at the same semantic scale***. By "same scale", we mean that all VG models overlook the hierarchical (or subevent) relations (Aldawsari and Finlayson, 2019; Yao et al., 2020) between these query sentences. For example, in Figure 1(c), the sentence "*Stir gently, leaving some small clumps of dry ingredients in the batter*" (S_2) is one of the subevents of "*Add the butter and milk to the mix*" (S_1), *i.e.*, S_1 and S_2 are at different semantic scales. Thus, the second assumption makes current VG models fail to perceive the semantic scales, and achieve unsatisfactory performance with multi-scale queries.

To this end, we propose a more realistic but challenging grounding task: **Article Grounding** (AG), which relaxes both above-mentioned assumptions. Specifically, given a video and a relevant article (*i.e.*, a sequence of sentences), AG requires the model to localize only "groundable" sentences to video segments, and these sentences are possibly at different semantic scales. To further avoid the manual annotations for the large-scale training set, in this paper, we consider a more meaningful setting: weakly-supervised AG (**WSAG**). As shown in Figure 2, the only supervision for WSAG is that the given video and article are about the same task[3].

Since there is no prior work on WSAG, we collect a new dataset, **YouwikiHow**, to benchmark the research. YouwikiHow is built on top of wikiHow articles and YouTube videos[4]. In particular, we group a wikiHow article and an arbitrary video about the same *task* as a document-level pair (cf. Figure 2). For the training set, we conduct a set of carefully designed operations to control the quality of training samples, *e.g.*, task filtering or sentence simplification. For the test set, we directly borrow the manual step grounding annotations in the existing CrossTask (Zhukov et al., 2019) dataset and propagate them to wikiHow article sentences.

In addition, we propose a simple but effective Dual loss constraint MIL-based method for WSAG, dubbed **DualMIL**. Specifically, **for the first assumption**, we relax the widely-used Multiple Instance Learning (MIL) loss into a two-level MIL loss. By "two-level", we mean that we regard *all sentences for each article* (sentence-level) and *all proposals for each sentence* (segment-level) as the "bag" at two different levels. Then, we obtain the global video-article matching score by aggregating all matching scores over the two-level bag. This two-level MIL inherently allows some queries that cannot be grounded in the video. Meanwhile, to avoid obtaining many highly-overlapping segments, we propose a single-sentence constraint to suppress the proposals whose neighbor proposals have higher matching scores with the query. **For the second assumption**, we enhance models' abilities in perceiving different semantic scale queries by considering these hierarchical relations across sentences. In particular, we assume that high-level sentences should be more likely to be grounded than its low-level sentences for highly matched proposals, and propose a cross-sentence constraint loss. We show the effectiveness of DualMIL over state-of-the-art methods through extensive ablations.

In summary, we make three contributions:

1. To the best of our knowledge, we are the first work to discuss the two unrealistic assumptions: all query sentences are groundable and all query sentences are at the same semantic scale. Meanwhile, we propose a meaningful WSAG task.
2. To benchmark the research, we collect the first WSAG dataset: YouwikiHow.
3. We further propose a simple but effective method DualMIL for WSAG, which consists of three different model-agnostic training objectives.

2 Related Work

Single Sentence & Multi-Sentence VG. Mainstream solutions for single sentence VG can be coarsely categorized into two groups: 1) *Top-down*

[3] A task means the same topic with clear and specific steps.
[4] https://www.wikihow.com/ & https://www.youtube.com/.

Methods (Hendricks et al., 2017; Gao et al., 2017; Zhang et al., 2019, 2020b; Chen et al., 2018; Yuan et al., 2019a, 2021; Wang et al., 2020; Xiao et al., 2021b,a; Liu et al., 2021b,a; Lan et al., 2022): They first cut given video into a set of segment proposals with different durations, and then calculate matching scores between query and all segment proposals. Their performance heavily relies on predefined rules for proposal settings (*e.g.*, temporal sizes). 2) *Bottom-up Methods* (Yuan et al., 2019b; Lu et al., 2019; Zeng et al., 2020; Chen et al., 2020a, 2018; Zhang et al., 2020a): They directly predict the two temporal boundaries of the target segment by regarding the query as a conditional input. Compared to their top-down counterpart, bottom-up methods always fail to consider the global context between two boundaries (*i.e.*, inside segment). In this paper, we follow the top-down framework and our DualMIL is model-agnostic.

Existing multi-sentence VG work all takes an assumption: the query sentences are ranked by their corresponding segments. This is an unrealistic and artificial setting. In contrast, real-world articles always do not meet this strict requirement, and most of the sentences are not even groundable in affiliated videos. In this paper, we take more realistic assumptions for the multi-sentence VG problem.

Weakly-Supervised VG. Since the agreements on the manually annotated target segments tend to be low (Otani et al., 2020), a surge of efforts aims to solve this challenging task in a weakly-supervised manner, *i.e.*, there are only video-level supervisions at the training stage. Currently, there are two typical frameworks: 1) *MIL-based* (Gao et al., 2019; Mithun et al., 2019; Chen et al., 2020b; Ma et al., 2020; Zhang et al., 2020c,d; Tan et al., 2021): They first calculate the matching scores between the query sentence and all segment proposals and then aggregate scores of multiple proposals as the score of whole "bag". State-of-the-art MIL-based methods usually focus on designing better positive/negative bag selections. 2) *Reconstruction-based* (Duan et al., 2018; Lin et al., 2020): They utilize the consistency between dual tasks *sentence localization* and *caption generation*, and infer the final grounding results from intermediate attention weights. Among them, the most related work to us is CRM (Huang et al., 2021), which considers both multi-sentence and weakly-supervised settings. Compared to CRM, our setting is more challenging: a) Sentences are from different scales; b) Not all sentences can be groundable; and c) Sentence sequences are not consistent with GTs.

Multi-Scale VL Benchmarks. With the development of large-scale annotation tools, hundreds of video-language (VL) datasets are proposed. To the best of our knowledge, three (types of) VL datasets also have considered the multiple semantic scale issue: 1) *TACoS Multi-Level* (Rohrbach et al., 2014): It provides three-level summaries for videos. In contrast, their middle-level sentences are more like extractive summarization (instead of abstractive). Thus, the grounding results for different-scale sentences may be the same. 2) *Movie-related* (Xiong et al., 2019; Huang et al., 2020; Bain et al., 2020): They always have multiple-level sentences to describe videos, such as overview, storyline, plot, and synopsis. They have two characteristics: a) Numerous sentences are abstract descriptions, *i.e.*, they do not have exact grounding temporal boundaries. b) The high-level summaries are more like highlights or salient events. 3) *COIN* (Tang et al., 2019): It defines multi-level predefined steps. Thus, it sacrifices the ability to ground any open-ended queries.

3 Dataset: YouwikiHow

We built **YouwikiHow** dataset from wikiHow articles and YouTube videos. As shown in Figure 2, we group a wikiHow article and any video about the same task as a pair. Thanks to the inherent hierarchical structure of wikiHow articles, we can easily obtain sentences from different scales: ***high-level*** summaries and ***low-level*** details. As in Figure 2, "*Pour the lemon juice into the pitcher.*" is a high-level sentence summary and "*You may add the pulp if .... along with the seeds.*" is a low-level sentence detail of this summary. In this section, we first introduce the details of dataset construction, and then compare YouwikiHow to existing VG benchmarks.

3.1 Dataset Construction

3.1.1 Training Set

Initial Visual Tasks. Each wikiHow article describes a sequence of steps to instruct humans to perform a certain "task", and these tasks range from physical world interactions to abstract mental well-being improvement. In YouwikiHow, we follow (Miech et al., 2019) and only focus on "visual tasks". This gives us 25K tasks to begin with.

Task-Related Videos. We also follow (Miech et al., 2019) and use the same preprocessing steps (*e.g.*, remove videos with few views or too short dura-

Dataset	#Videos	Avg Sents per Video	#Tasks	Multi-Moment per Query	Open Vocabulary	Support Multi-Scale	May Not Groundable
DiDeMo (Hendricks et al., 2017)	10.6K	3.9	—	—	✓	—	—
Charades-STA (Gao et al., 2017)	6.7K	2.4	—	—	✓	—	—
ANet-Caps (Krishna et al., 2017)	15K	4.8	—	—	✓	—	—
YouCook2 (Zhou et al., 2018)	2K	7.7	89	✓	—	—	—
TVR (Lei et al., 2020)	21.8K	5.0	—	—	✓	—	—
QVHighlights (Lei et al., 2021)	10.2K	1.0	—	✓	✓	—	—
CrossTask (Zhukov et al., 2019)	4.7K	7.4∼8.8	83	✓	—	—	—
COIN (Tang et al., 2019)	11.8K	3.9	180	—	—	✓	—
YouwikiHow (training set)	47K	20.8	1,398	✓	✓	✓	✓

Table 1: Comparison between YouwikiHow and other prevalent video grounding or step segmentation benchmarks.

tions) to obtain initial task-related videos for each task. To further control the quality and ensure sufficient training videos for each task, we restrict the videos to top 50 search results, and the number of training videos for each task to be at least 30. This step prunes the number of tasks from 25K to 2.3K.

Sentence Quality Control. Firstly, to avoid over-long articles, we filter out all the tasks with verbose sentences. Specifically, we set the max number of sentence summaries and details to 10 and 30, respectively. This filtering step decreases the task number to 1.4K. Meanwhile, since original wikiHow articles usually contain unimportant modifiers or quantifiers, we further conduct rule-based *sentence simplification* (Al-Thanyyan and Azmi, 2021) based on POS and dependency parse tags[5].

3.1.2 Test Set

For the test set, we directly build on top of the existing CrossTask (Zhukov et al., 2019) and reuse their manual temporal grounding annotations. Specifically, CrossTask is originally proposed for step segmentation, which consists of 18 primary wikiHow tasks. For each task, it collects corresponding YouTube videos and annotates the temporal grounding boundaries for each video corresponding to the predefined task-specific steps. Then, we manually link the step to the wikiHow articles[6] and propagate these annotations as the ground-truth for wikiHow sentences. We conduct the same sentence simplification steps on all the wikiHow articles in the test set, and remove the task with over-long articles[7]. Unfortunately, when we perform manually linking between CrossTask steps and wikiHow articles, we found it is difficult to link these steps to low-level details and almost all steps are linked to high-level summaries. To this end, we further design different evaluation metrics for high-/low- level sentences to bypass these limitations. (Details are in Sec. 5.1.)

3.2 Comparison with Existing VG Datasets

We compare our collected YouwikiHow with other prevalent VG or step segmentation datasets in Table 1. In the training set, we have a total of 1,398 wikiHow tasks, and each task has an average of 33.88 videos. For each task, there are 6.01 high-level sentence summaries and 14.79 low-level sentence details. Compared to existing VG datasets, YouwikiHow has more training videos (47K vs. 21.8K), and much more query sentences for each video (20.8 vs. 7.7). More importantly, it supports multi-scale queries and the query sentences may not be grounded in the video. Compared to step segmentation datasets, it not only has much more queries for each video and more diverse training tasks (1,398 vs. 180), but also supports both open-vocabulary queries and multi-scale queries.

4 Proposed Approach for WSAG

Problem Formulation. WSAG is defined as follows: Given an untrimmed video V and a relevant

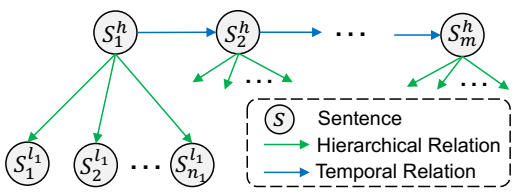

Figure 3: Illustration of the multi-scale structures of A.

[5]For example, given the original sentence "*Then, mix in 1 teaspoon (4.9 mL) of vanilla extract, followed by 1 teaspoon (2.6 grams) of cinnamon.*", sentence simplification can prune these unimportant modifier (gray words) and obtain a new sentence: "*Then, mix in vanilla extract, followed by cinnamon.*"

[6]For example, we can easily link CrossTask steps "brake on" (*Change a Tire*) or "attach shelve" (*Build Simple Floating Shelves*) to the sentence "*Apply the parking brake and put car into 'Park' position.*" or "*Attach the shelf mount to the wall*" in their corresponding wikiHow articles, respectively.

[7]We remove three tasks: *Make Kimichi Fried Rice*, *Add Oil to Your Car*, and *Make French Strawberry Cake*. All these tasks have over 60 sentences in their wikiHow articles.

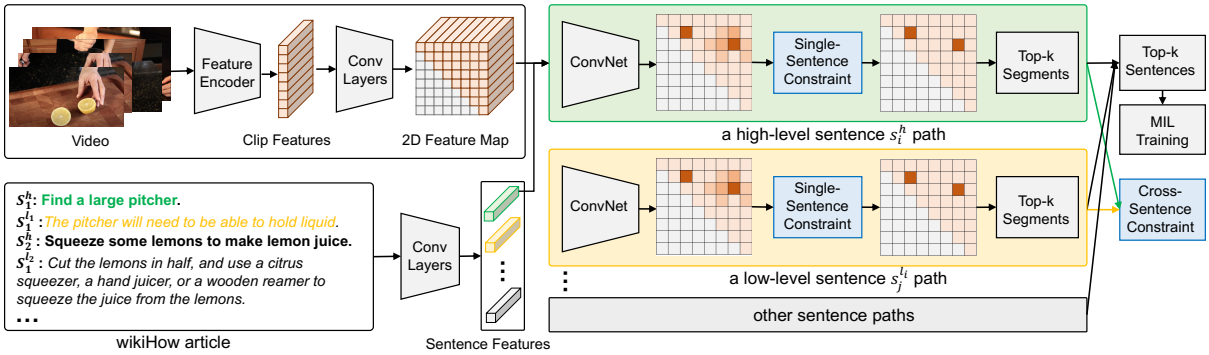

Figure 4: The overview of the article grounding architecture with the proposed DualMIL.

article A with multi-scale sentences, WSAG needs to predict all possible temporal locations for all groundable sentences, *i.e.*, one sentence may refer to either multiple segments or even none.

In this paper, we consider sentences at two scales. Specifically, as shown in Figure 3, article A is organized as $A = \{s_1^h, s_1^{l_1}, ..., s_{n_1}^{l_1}; s_2^h, ...; s_m^h, ..., s_{n_m}^{l_m}\}$, where s_k^h is the k-th high-level summary, and $s_i^{l_k}$ is the i-th low-level details of s_k^h. There are m high-level summaries in total, and each high-level summary s_k^h has n_k low-level details. To show more generalized abilities, in test stage, we assume that we do not know the scale prior of each sentence.

In this section, we first go through the architecture for grounding in Sec 4.1. Then, we detail each component of DualMIL in Sec 4.2.

4.1 Basic Visual Grounding Architecture

Since DualMIL is a model-agnostic training strategy, we follow a SOTA proposal-based model 2D-TAN (Zhang et al., 2020b) and use it as our baseline. As shown in Figure 4, it consists of three parts:

Video Feature Encoding. Given video V, we first use a pretrained video feature extractor to extract clip features, and sample the video features evenly to N clips. Then, we utilize the 2D-map proposal strategy: All the segment proposals can be organized into a 2D temporal map M, and each element $m_{ij} \in M$ represents the candidate segment which starts from clip$_i$ and ends at clip$_j$. We extract each proposal feature by averaging all inside clip features, and then stack a few conv-layers to further encode the context. Finally, we obtain 2D feature map $\boldsymbol{F}^M \in \mathbb{R}^{N \times N \times d_v}$, and each element $\boldsymbol{F}_{ij}^M$ denotes the feature of segment proposal m_{ij}.

Text Feature Encoding. For each sentence $s_i = \{w_j^i\}$ in article A, we first use the GloVe embedding (Pennington et al., 2014) to encode each word w, and then feed all word embeddings into a Bi-LSTM. The final hidden state of Bi-LSTM is taken as the feature of sentence, denoted as $\boldsymbol{F}^{S_i} \in \mathbb{R}^{d_s}$.

Multimodal Matching. After obtaining the video feature $\boldsymbol{F}^M$ and all sentence features $\{\boldsymbol{F}^{S_i}\}$, we then fuse these two features by Hadamard product:

$$\tilde{\boldsymbol{F}}_{ij,k} = \boldsymbol{w}_s \boldsymbol{F}^{S_k} \odot \boldsymbol{w}_v \boldsymbol{F}_{ij}^M, \quad (1)$$

where $\boldsymbol{w}_s \in \mathbb{R}^{d_h \times d_s}$ and $\boldsymbol{w}_v \in \mathbb{R}^{d_h \times d_v}$ are two learnable MLPs, which map two modality features into a common space. Reorganizing the $\tilde{\boldsymbol{F}}_{ij,k}$ into the 2D map format, we can obtain $\tilde{\boldsymbol{F}}_k \in \mathbb{R}^{N \times N \times d_h}$, which denotes the fused feature between sentence s_k and all segment proposals M.

Later, we adopt several conv-layers to obtain context-aware multimodal 2D feature maps. And these feature maps are fed into the classifier to predict all the matching score maps $\{\boldsymbol{P}^k\}$, where $\boldsymbol{P}^k \in \mathbb{R}^{N \times N}$ denotes the matching scores between all segment proposals M and sentence s_k.

4.2 DualMIL Training Objectives & Inference

4.2.1 Two-level MIL Training Objective

Since not all sentences in article A are groundable to the given video V, we only select the top-k_1 sentences with the highest matching scores to represent the whole article. As for the matching score between each sentence and the video, we average the similarity scores among the top-k_2 proposals:

$$sim(V, s_k) = avg\{\text{top-}k_2 \max_{ij} \boldsymbol{P}_{ij}^k\}, \quad (2)$$

where $sim(V, s_k)$ denotes the similarity score between video V and sentence s_k. Similarly, we use $sim(V, A)_i$ to denote the similarity score between the top-i sentence in A with video V (*i.e.*, $i \leq k_1$).

We train the whole model with the ranking loss. Specifically, we treat video V and its same-task article A as *positive* pair (V, A). Then we randomly

replace the video or article with other-task videos or articles to obtain *negative* pairs, denoted as (V^-, A) and (V, A^-) respectively. Then, the two-level MIL loss is written as $\mathcal{L}_{\text{MIL}} = \sum_i \sum_j \mathcal{L}^{ij}_{\text{MIL}}$, and

$$\begin{aligned}\mathcal{L}^{ij}_{\text{MIL}} = &\max(0, \Delta - \text{sim}(V, A)_i + \text{sim}(V^-, A)_j) \\ &+ \max(0, \Delta - \text{sim}(V, A)_i + \text{sim}(V, A^-)_j),\end{aligned} \tag{3}$$

where Δ is a predefined margin.

4.2.2 Single-Sentence Constraint

Since each sentence may be grounded in multiple segments, we need to predict the similarity scores between each query sentence and all segment proposals. To force WSAG models to make sparse predictions, we propose the single-sentence constraint to enhance the two-level MIL training. By "sparse", we mean that only a few proposals are selected as results for each groundable sentence.

Specifically, before selecting the top-k_2 segment proposals as in Eq. (2), we conduct a sparse filtering step to suppress (or filter out) the proposals by two rules: 1) In local highly-overlapped neighbors, there are other proposals with higher video-sentence matching scores. 2) The matching score is much less than the proposal with the highest score.

From an implementation perspective, we can use a simple max-pooling layer with kernel size K and a threshold δ to realize single-sentence constraint. Then, we can obtain a new filtered $\tilde{\boldsymbol{P}}^k$, and calculate a similar MIL loss with $\tilde{\boldsymbol{P}}^k$ following Eq. (2) and Eq. (3). (Ablations on K and δ are in Sec. 5).

Highlights. Compared to the existing constraint strategy by selecting the proposal with the highest score as extra pseudo GT (Wang et al., 2021), our solution avoids selecting unstable pseudo GT (*i.e.*, more robust), and it is more suitable for the setting of any number of GT segments for each query.

4.2.3 Cross-Sentence Constraint

To force WSAG models to perceive multi-scale queries, we propose the cross-sentence constraint. Specifically, we assume that high-level sentences should be more like to grounded than its low-level sentences for highly matched proposals. The reason is that today's multimodal coreference relations between query sentence and GT video segment discussed in grounding works contain both "*identical*" and "*hierarchical*" relations. Let's take two extreme cases as examples: 1) If the low-level sentence is identical to the video segment proposal, then its high-level sentence is also coreference to the proposal (hierarchical relation). 2) If the high-level sentence is identical to the proposal, then its low-level sentence is only partially coreference to the proposal. Thus, we propose the cross-sentence constraint by limiting the proposal matching scores between a high-level and low-level sentence pair.

Obviously, if the proposal itself is unrelated to the low-level sentence, this constraint is meaningless. Thus, we use the low-level sentence matching score as the loss weight, and the contraint loss is:

$$\mathcal{L}_{\text{CS}} = \sum_{h=1}^{m} \sum_{k=1}^{n_h} \sum_{ij} \max(0, \alpha - \boldsymbol{P}^h_{ij} + \boldsymbol{P}^{l_h,k}_{ij}) \cdot \boldsymbol{P}^{l_h,k}_{ij}, \tag{4}$$

where $\boldsymbol{P}^h_{ij}$ is the matching score between h-th high-level sentence and proposal m_{ij}, and $\boldsymbol{P}^{l_h,k}_{ij}$ is the matching score between k-th low-level sentence of s^h_h and proposal m_{ij}. α is a predefined margin, and the impact of α is discussed in Table 3.

Highlights. Since multimodal hierarchical relation is always difficult to predict, the main effect of the cross-sentence constraint is to avoid the case: a low-level sentence has a high matching score with the proposal while its high-level summary is not.

4.2.4 Inference

In the test stage, given a video and a relevant article, we first predict the matching scores between each sentence and all proposals, and then we conduct non-maximum suppression (NMS) to filter out the proposals with highly overlaps but smaller scores. Then, we can simply combine all predictions from different sentences based on their matching scores.

To further consider the semantic relations between sentences at test stage, we use a **Structure-NMS**, inspired by Soft-NMS (Bodla et al., 2017), to suppress the segments which violate structure constraints. More details are left in the appendix.

5 Experiments

5.1 Experimental Settings

Metrics. We used *Recall@K* (**R@K**) over different IoU thresholds (0.1/0.3/0.5) to evaluate each video-article pair. Specifically, we ranked all segment-sentence pairs based on their matching scores, and calculated the recalls of all GT annotations within top-K predictions[8]. A prediction is hit if its IoU with GT is larger than the threshold. Since we only

[8]We used Recall as metrics for two reasons: 1) Following existing VG works, they also use Recall@K as the main metric. 2) Due to our GT propagation rules for the test set, we may miss some GT annotations, *i.e.*, (cf. Sec. Limitations).

Model	R@50 (IoU)			R@100 (IoU)		
	0.1	0.3	0.5	0.1	0.3	0.5
Baseline	26.60	14.98	6.48	44.05	24.81	10.82
Baseline w/ *Single-Sentence Constraint*						
K=7, δ=0.9	27.86	16.00	7.00	41.76	23.95	10.63
K=7, δ=0.7	30.02	**17.38**	**7.70**	43.50	25.50	11.68
K=7, δ=0.5	**30.37**	17.30	7.51	**47.57**	**27.21**	**11.97**
K=5, δ=0.5	27.36	15.15	6.67	45.24	25.62	11.37
K=3, δ=0.5	26.25	14.66	6.66	45.99	25.95	11.49

Table 2: Ablations (%) on single-sentence constraint.

Model	R@50 (IoU)			R@100 (IoU)		
	0.1	0.3	0.5	0.1	0.3	0.5
Baseline	26.60	14.98	6.48	44.05	24.81	10.82
Baseline w/ *Cross-Sentence Constraint*						
α = 0.1	33.79	18.98	8.21	52.60	29.37	12.83
α = 0.0	**34.73**	**19.23**	**8.46**	**55.07**	**30.66**	**13.46**
α = -0.1	32.63	18.38	8.05	52.28	29.42	12.93

Table 3: Ablations (%) on cross-sentence constraint.

have GT annotations for high-level summaries, we also proposed *Recall@K meet Constraint* (**RC@K**) as a supplementary metric for low-level sentences. Since we assume the temporal grounding results of low-level sentences should be inside its high-level manual annotations, we calculated the percentages of low-level sentence predictions that meet the constraint. Note that RC@K is not strictly accurate.

Implementation Details. Given a reference video V, we used a pretrained S3D extractor (Miech et al., 2020) to extract initial clip features. The number of initial clips was set to 256. For text sentences, following prior VG works, we truncated or padded each sentence to a maximum length of 25 words. In the training stage, to save GPU memory, we randomly sample 20 sentences if the articles have more than 20 sentences. All the dimensions of the hidden features were set to 512. In the multimodal matching, we used a three-layer convolutional network to encode context. Its kernel size and strides were set to 3 and 1, respectively. We trained the whole network with Adam optimizer for 100 epochs. The initial learning rate was set to 0.0001, and the batch size was set to 32. The loss weights of two-level MIL loss (for both models with and without single-sentence constraint) and cross-sentence constraint loss were set to 1.0 and 0.1, respectively. The predefined margin Δ for MIL training was set to 0.3. For the model with cross-sentence constraint, to ensure the predicted low-level sentence matching scores are reliable, we first train the model with MIL loss solely at a warm-up stage, and then add cross-sentence constraint loss for further training.

Strategies			R@50 (IoU)			R@100 (IoU)		
SS	CS	NMS	0.1	0.3	0.5	0.1	0.3	0.5
✗	✗	✗	26.60	14.98	6.48	44.05	24.81	10.82
✗	✗	✓	26.92	15.22	6.63	44.14	24.87	10.84
✓	✗	✗	30.37	17.30	7.51	47.57	27.21	11.97
✗	✓	✗	34.73	19.23	8.46	**55.07**	30.66	13.46
✓	✓	✓	**40.21**	**22.98**	**9.99**	54.55	**31.28**	**13.96**

Table 4: Ablations (%) on the effectiveness of each part, where "SS", "CS", and "NMS" denote single-/cross-sentence constraint and structure-NMS, respectively.

5.2 Ablation Studies

We run a number of ablations to analyze the impact of different hyperparameters of each component, and the effectiveness of each component.

Ablation on Single-Sentence Constraint. The impacts of two hyperparameters in the single-sentence constraint (*i.e.*, kernel sizes K and thresholds δ) are reported in Table 2. From the results, we can observe that: 1) For most hyperparameter settings, the single-sentence constraint can consistently improve models' performance. 2) The Model with setting $K = 7$ and $\delta = 0.5$ achieves the best results.

Ablation on Cross-Sentence Constraint. The impact of different margin α in the cross-sentence constraint are reported in Table 3. From results, we can observe that the performance gains are robust to different α, and the model with $\alpha = 0$ achieves the best performance. It is worth noting that a negative α (relaxed constraint) is still effective, which proves the claimed main effects in Sec. 4.2.3.

Ablation on Structure-NMS. The results of the models with and without structure-NMS are illustrated in Figure 5. From the results, we can observe that structure-NMS can significantly improve the performance of tasks with high agreements (*e.g.*, *Change a Tire*, or *Grill Steak*). In contrast, it may hurt the performance of tasks with low agreements.

Effectiveness of Each Strategy. The ablation studies on each strategy are reported in Table 4. From Table 4, we have the following observations: 1) Compared to the baseline, each strategy can consistently improve performance on both R@50 and R@100 metrics. 2) The full model achieves the best R@50 and R@100 over different IoUs.

5.3 Comparisons with State-of-the-Art

Baselines. We compared our proposed DualMIL with a set of state-of-the-art baselines. Specifically, we investigated three types of baselines:

Type1: State-of-the-art WSVG models. We com-

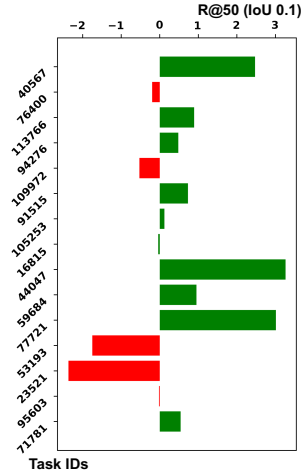

Figure 5: Performance gains (%) between models w/ & w/o structure-NMS. Task ids are ranked by the agreement between the order of groundable sentences and GT segments (cf. Appendix).

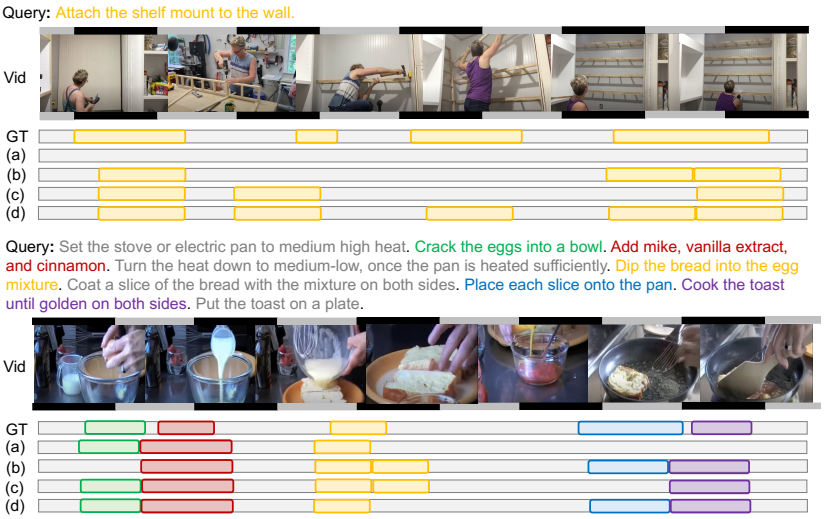

Figure 6: **Upper**: An example of a query with multiple GT segments. **Below**: Given a video and an article (only high-level sentences), GT segments of all groundable sentences are shown (with corresponding colors). (a) - (d) denotes baseline, baseline w/ single-sentence const., baseline w/ cross-sentence const., and full model, respectively. Top-50 predictions overlapped with GT are shown.

	Model	MM Pretrain	**R@50** (IoU)			**R@100** (IoU)			RC@50 (IoU)		
			0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Type1	RandomGuess*		19.55	5.22	1.67	33.05	10.46	3.88	7.23	1.87	0.66
	MIL-XE ($\text{WSTAN}_{\text{base}}$)		26.61	15.64	7.24	36.19	20.68	9.75	13.53	10.31	8.95
	WSTAN†		16.80	2.39	0.50	16.80	2.39	0.50	6.87	0.81	0.19
Type2	MIL-NCE-max	✓	33.29	11.93	4.84	39.54	14.11	5.76	11.69	3.11	1.07
	MIL-NCE-avg	✓	**42.81**	**23.96**	**12.75**	**56.62**	**31.66**	**16.79**	16.25	7.62	3.82
Type3	MIL-NCE+$\text{WSTAN}_{\text{base}}$ a	✓	28.70	12.57	5.82	42.50	18.97	9.06	11.00	4.04	1.53
	MIL-NCE+$\text{WSTAN}_{\text{base}}$ b	✓	35.10	15.53	7.16	48.66	22.02	10.74	9.73	3.81	1.84
	MIL-NCE+$\text{WSTAN}_{\text{base}}$ c	✓	33.16	18.53	8.83	51.28	29.31	14.28	14.23	10.87	9.28
	DualMIL (Ours)		40.21	22.98	9.99	54.55	31.28	13.96	11.31	8.35	7.07

Table 5: Performance (%) comparison with SOTA baselines. All listed methods use the same proposal settings. "MM Pretrain" denotes these models use large-scale multimodal pretraining features. * results are averaged by five different random seeds. Model a/b/c in Type3 denotes model with different thresholds. † denotes reimplementation results using official codes. The **best** and second best results are denotes with corresponding formats.

pared with **WSTAN** (Wang et al., 2021), which builds on top of a cross-entropy (XE) based MIL backbone. For completeness, we also reported results of the WSTAN backbone (dubbed **MIL-XE**), and a random guess (**RandomGuess**) baseline.

Type2: Pretrained multimodal video-text retrieval models (*e.g.*, **MIL-NCE** (Miech et al., 2020)). We show the *zero-shot* results of two variants by max-pooling or average-pooling the clip features inside the boundaries of video segment proposals.

Type3: Two-stage model. Since today's WSVG models assume all the sentences can be grounded to the video, a straightforward two-stage solution is: Using pretrained video-text retrieval models to select all groundable sentences first, and then training a WSVG model with selected sentences. Obviously, we need to manually set a threshold to filter out sentences at the first stage, we reported results of three variants with different thresholds.

Results. All results are reported in Table 5. From Table 5, we have the following observations: 1) For Type1 methods, the simple baseline MIL-XE can achieve good performance. However, the SOTA model WSTAN with other more advanced designs only performs similarly with RandomGuess, which proves existing SOTA WSVG models fail to work in these more realistic settings. 2) For Type2 methods, the performance gaps between different pooling operations are large. Although these large-scale pretrained models can achieve exemplary zero-shot performance, they are not robust enough and heavily rely on different heuristic rules. 3) For Type3

methods, the model with different thresholds also behavior differently, *i.e.*, these two-stage methods are not robust either. 4) In contrast, our proposed DudalMIL can achieve satisfactory performance with relatively consistent gains.

5.4 Visualizations

We illustrated two examples in Figure 6. For the first example, we only show the grounding results of one query sentence (from article "*Build Simple Floating Shelves*") with multiple ground-truth segments. For the second example, we show the grounding results of all high-level sentences of the article ("*Make French Toast*"). From Figure 6, we observe that: Both the proposed single-sentence constraint and cross-sentence constraint can help to ground some missing segments in top-K predictions. Meanwhile, both constraints are complementary, *i.e.*, the full model achieves the best results.

6 Conclusions

In this paper, we discussed the weaknesses of default assumptions in existing video grounding work, and proposed a more challenging task: weakly-supervised article grounding (WSAG). To facilitate the research in this direction, we collected the first WSAG dataset YouwikiHow. Further, we proposed DualMIL for WSAG, including a two-level MIL loss and a single-/cross-sentence constraint loss. This work paves the way for a number of exciting future works: 1) designing more reasonable backbones for multiple sentence inputs by considering their semantic relations; 2) extending to more general domains beyond instructional articles.

Acknowledgments

We thank the anonymous reviewers' helpful suggestions. This research is based upon work supported by U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Limitations

The main limitations of this work are about the collected dataset *YouwikiHow*. Specifically, we can discuss them from the two following aspects:

Dataset Creation. Since we focus on WSAG, the manner of creating the training set of YouwikiHow is acceptable. However, to save the manual annotations for the test set, we only propagate the annotations from the existing CrossTask (Zhukov et al., 2019) dataset. Although this solution is much cheaper, it introduces two types of potential errors in the "ground-truth" annotations for evaluation: 1) When manually mapping the "step" in CrossTask to the "sentence" in the wikiHow article, we found it not always be one-to-one perfect mapping. In a few cases, multiple sentences may refer to a single step or multiple steps may refer to a single sentence. Thus, the original ground-truth annotations for each CrossTask step may not be exactly accurate for its mapped sentence regarding the same video. 2) Since each wikiHow article has much more sentence queries than original step queries in CrossTask, many wikiHow sentences cannot be mapped to these predefined steps, *i.e.*, these wikiHow sentences will not have any "ground-truth" annotations. However, these sentences may be groundable in some specific videos.

Domain Coverage. Since we obtain the explicit multi-scale sentences from the inherent hierarchical structures of wikiHow articles, these wikiHow articles are mainly about instructional articles. Thus, the main domain of our YouwikiHow dataset focuses on instructional articles/videos, *i.e.*, the model trained in our dataset may suffer from performance drops when they are applied to other domain daily multimodal assets.

For the first limitation, we mitigate its impact by using more relaxed metrics: R@K or RC@K. Of course, the most accurate solution is checking all the annotations between any video-article pairs.

Ethics Statement

The proposed dataset and method aim to improve the performance of temporal grounding models in more realistic settings. Advancements in visual grounding help the deployment of visual grounding (or article grounding) models in our daily applications. Since we mainly focus on the two unrealistic assumptions in existing grounding models, our work does not introduce new ethical concerns. The only potential ethical concern is that any language-query based applications run the risk of using biased or offensive words (or descriptions) — video grounding is no exception. In the future, we can try to incorporate a preprocessing step to avoid or correct biased or offensive content.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, pages 1–36.

Mohammed Aldawsari and Mark A Finlayson. 2019. Detecting subevents using discourse and narrative features. In *ACL*.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*.

Peijun Bao, Qian Zheng, and Yadong Mu. 2021. Dense events grounding in video. In *AAAI*, pages 920–928.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-nms–improving object detection with one line of code. In *ICCV*, pages 5561–5569.

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *EMNLP*.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020a. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, pages 10551–10558.

Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020b. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. In *arXiv*.

Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *NeurIPS*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275.

Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. Wslln: Weakly supervised natural language localization networks. In *EMNLP*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, pages 7199–7208.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *ECCV*, pages 709–727.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*, pages 706–715.

Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. 2022. A closer look at debiased temporal sentence grounding in videos: Dataset, metric, and approach. *ACM Trans. Multimedia Comput. Commun. Appl.*

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, pages 447–463.

Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, pages 11539–11546.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2021a. Adaptive proposal generation network for temporal sentence localization in videos. In *EMNLP*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021b. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*.

Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, pages 5147–5156.

Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, pages 156–171.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *GCPR*, pages 184–195.

Fengyuan Shi, Limin Wang, and Weilin Huang. 2021. End-to-end dense video grounding via parallel regression. *arXiv*.

Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, pages 2083–2092.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216.

Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, pages 12168–12175.

Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2021. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*.

Shaoning Xiao, Long Chen, Kaifeng Gao, Zhao Wang, Yi Yang, and Jun Xiao. 2022. Rethinking multimodal alignment in video question answering from feature and sample perspectives. In *EMNLP*.

Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. 2021a. Natural language video localization with learnable moment proposals. In *EMMLP*.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021b. Boundary proposal network for two-stage natural language video localization. In *AAAI*.

Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. 2019. A graph-based framework to bridge movies and synopses. In *ICCV*, pages 4592–4601.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly supervised subevent knowledge acquisition. In *EMNLP*.

Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *SIGIR*, pages 829–832.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. 2021. A closer look at temporal sentence grounding in videos: Dataset and metric. In *ACM MM*, pages 13–21.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019a. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019b. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, pages 9159–9166.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *CVPR*, pages 10287–10296.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *ACL*.

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

Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020c. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020d. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *NeurIPS*, pages 18123–18134.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*, pages 3537–3545.

Appendix

The appendix is organized as follows:

- More details about the structure-NMS are in Sec. A.
- More experimental details are in Sec. B.
- More ablation studies are in Sec. C.
- The statistics about the agreement of GT temporal orders are in Sec. D.

	Setting	R@50			R@100		
		0.1	0.3	0.5	0.1	0.3	0.5
Prediction	N=8	26.57	10.69	4.13	43.55	17.39	6.73
	N=12	26.01	13.46	4.64	**44.15**	21.61	9.01
	N=16	**26.60**	**14.98**	**6.48**	44.05	24.81	10.82
	N=24	24.69	14.81	6.24	39.79	**25.25**	**11.29**
GT	N=8	64.14	26.88	11.57	66.82	28.10	12.29
	N=12	68.02	35.31	15.77	74.06	38.24	17.52
	N=16	**71.79**	**42.47**	**19.90**	**80.86**	47.88	23.00
	N=24	62.45	41.07	19.29	76.05	**50.43**	**24.03**

Table 6: Ablations on different proposal settings. "GT" denotes the results with only groundable sentences.

A More Details about Structure-NMS

Given all detected segments for each groundable sentence in the article, we hope these segments themselves also meet the same semantic relations as their query sentences (temporal or hierarchical relations). Since we assume that we do not know the scale prior of each sentence at the test stage, currently we only consider the temporal relations.

More specifically, given two query sentences s_i and s_j. If s_i appears earlier than s_j in the corresponding article, we hope the grounding segments for s_i should be earlier than s_j too. Following Soft-NMS (Bodla et al., 2017), we also multiple a coefficient to decrease the matching score of the proposals which violate this temporal constraint, and the coefficient is proportional to their IoU. Let's take a concrete example. If the predicted segments for s_i and s_j are $[l_s^i, l_e^i]$ and $[l_s^j, l_e^j]$, and their matching scores are p^i and p^j ($p^i < p^j$). After selecting the segment $[l_s^j, l_e^j]$ into top-K predictions, we then slightly decrease the matching score p^i by:

$$IoU_{bad} = \frac{\max(l_s^j - l_s^i, 0) + \max(l_e^i - l_e^j, 0)}{\max(l_e^i, l_e^j) - \min(l_s^i, l_s^j)}, \quad (5)$$
$$p_{new}^i = p^i * \exp(-(IoU_{bad}) * *2/\text{const}),$$

where const is a constant number.

B More Experimental Details

More Details about RC@K. Since we hope the grounding segment of the low-level sentence is inside that of their high-level summary, we calculate RC@K the same way as plain recall with only one exception: if the low-level prediction is totally inside their high-level ground-truth annotations, the prediction is regarded as hit regardless of the IoU.

ID	Task Name	Agree.
40567	Change a Tire	96.94%
76400	Make French Toast	93.11%
113766	Grill Steak	89.33%
94276	Make Meringue	86.71%
109972	Make Banana Ice Cream	86.89%
91515	Make Pancake	85.49%
105253	Make Bread and Butter Pickles	83.61%
16815	Jack Up a Car	78.01%
44047	Make Lemonade	77.67%
59684	Build Simple Floating Shelves	75.39%
77721	Make Irish Coffee	71.21%
53193	Make a Latte	69.64%
23521	Make Jello Shots	68.61%
95603	Make Kerala Fish Curry	62.75%
71781	Make Taco Salad	54.39%

Table 7: The statistics about the agreement between the order of ground-truth query sentence and the order of their corresponding ground-truth segments.

C More Ablation Studies

Impact of Proposal Settings. For proposal-based VG methods, a notorious weakness is that their performance is heavily affected by different proposal settings. To this end, we explored the impact of different proposal settings in our baseline framework, and the results are reported in Table 6. From Table 6, we can observe that the model achieves the best performance in most metrics when N is 16. The performance gap between the "prediction" and "GT" settings also shows that the main bottleneck for current article grounding models is detecting groundable sentences for the video-article pair.

D Statistics about the Agreement of GT Temporal Orders

The agreement between the order of all groundable sentences and the order of their corresponding ground-truth segments of the test set are reported in Table 7.