# End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding

**Mengze Li[1], Tianbao Wang[1], Haoyu Zhang[1], Shengyu Zhang[1], Zhou Zhao[2,3*],**
**Jiaxu Miao[1*], Wenqiao Zhang[1*], Wenming Tan[4], Jin Wang[4], Peng Wang[5],**
**Shiliang Pu[4], Fei Wu[2,3*]**

[1]Zhejiang University [2]Shanghai Institute for Advanced Study of Zhejiang University
[3]Shanghai AI Laboratory [4]Hikvision Research Institute
[5]Northwestern Polytechnical University

## Abstract

Natural language spatial video grounding aims to detect the relevant objects in video frames with descriptive sentences as the query. In spite of the great advances, most existing methods rely on dense video frame annotations, which require a tremendous amount of human effort. To achieve effective grounding under a limited annotation budget, we investigate one-shot video grounding, and learn to ground natural language in all video frames with solely one frame labeled, in an end-to-end manner. One major challenge of end-to-end one-shot video grounding is the existence of videos frames that are either irrelevant to the language query or the labeled frames. Another challenge relates to the limited supervision, which might result in ineffective representation learning. To address these challenges, we designed an end-to-end model via Information Tree for One-Shot video grounding (IT-OS). Its key module, the information tree, can eliminate the interference of irrelevant frames based on branch search and branch cropping techniques. In addition, several self-supervised tasks are proposed based on the information tree to improve the representation learning under insufficient labeling. Experiments on the benchmark dataset demonstrate the effectiveness of our model.

## 1 Introduction

Natural language spatial video grounding is a vital task for video-text understanding (Luo and Shakhnarovich, 2017; Zhou et al., 2019; Hu et al., 2019; Zhang et al., 2020b; Li et al., 2021), which aims to detect the objects described by the natural
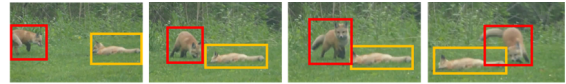


Caption A: A fox from the distance is coming close to another one on the ground.
Caption B: A fox is lying on the grass and watching another one coming close.

Figure 1: An example of spatially grounding natural language in video frames.

language query from each video frame, as shown in Figure 1. There is a substantial and rapidly-growing research literature studying this problem with dense annotations (Li et al., 2017; Yamaguchi et al., 2017; Sadhu et al., 2020), where each frame that contains objects relevant to the language query will be manually labeled with bounding boxes. Obviously, such annotations require tremendous human effort and can hardly be satisfied in real-world scenarios. Recently, some works have investigated weakly-supervised video grounding with solely the video-text correspondence rather than object-text annotations (Huang et al., 2018; Chen et al., 2019a; Shi et al., 2019; Chen et al., 2019b; Zhou et al., 2018). However, the performance is less satisfied with such weak supervision. In practice, we are more likely to have a limited annotation budget rather than full annotation or no annotation. In addition, as humans, after experiencing the language query and one frame object paired together for the first time, we have the ability to generalize this finding and identify objects from more frames. Towards this end, we investigate another practical problem setting, *i.e.*, one-shot spatial video grounding, with solely one relevant frame in the video labeled with bounding boxes per video.

Existing methods that are devised for supervised video grounding are not directly applicable to this novel setting. We summarize several critical challenges:

• On the one hand, most of them incorporate a multi-stage training process, *i.e.*, firstly training a clip localization module, and training an object

First author.
mengzeli@zju.edu.cn

*Corresponding author.
zhaozhou@zju.edu.cn
jiaxu.miao@yahoo
wenqiaozhang@zju.edu.cn
wufei@zju.edu.cn

8707

localization module in the second stage. However, in one-shot spatial video grounding, there are no temporal annotations, which indicate the start/end time of the relevant clip, to train the clip localization module. Moreover, many of them extract video features in a pre-processed manner using feature extractor or object detector pretrained on large-scale datasets. However, independent modeling limits the cooperation of different modules, especially when the labels are few. Therefore, it is in urgent need to derive an end-to-end training framework for one-shot spatial video grounding.

- On the other hand, there are video frames that are either irrelevant to the natural language query or the labeled frames. These irrelevant frames might increase the computation complexity of end-to-end training, and bring confounding between the frame label and (irrelevant) visual features.

- Lastly, with fewer supervision signals, deep representation learning might become error-prone or easily under-fitting, especially for end-to-end training.

To address these challenges, we devise an end-to-end model via the Information Tree for the One Shot natural language spatial video grounding (IT-OS). Different from previous works, we design a novel tree structure to shield off the one-shot learning from frames that are irrelevant to either the language query or the labeled frame. We devise several self-supervised tasks based on the tree structure to strengthen the representation learning under limited supervision signals. Specifically, the calculation processes of the key module, information tree, contains four steps: (1) To construct the information tree, we view video frame features as nodes, and then compress the adjacent nodes to non-leaf nodes based on the visual similarity of themselves and the semantic similarity with the language query; (2) We search the information tree and select branch paths that are consistently relevant to the language query both in the abstractive non-leaf node level and in the fine-grained leaf node level; (3) We drop I) the leaf nodes that do not belong the same semantic unit with the labeled node; and II) the non-leaf nodes on the low relevance branch paths. We also down-weight the importance of the leaf nodes that belong to the same semantic unit with the labeled node but are on the low relevance paths; (4) Finally, we input

the extracted and weighted information to the transformer, and conduct training with the one-shot label and self-supervised tasks, including masked feature prediction and video-text matching. We note that both the information tree and the transformer are jointly trained in an end-to-end manner.

We conduct experiments on two benchmark datasets, which demonstrate the effectiveness of IT-OS over state-of-the-arts. Extensive analysis including ablation studies and case studies jointly demonstrate the merits of IT-OS on one-shot video grounding. Our contributions can be summarized as follows:

- To the best of our knowledge, we take the initiative to investigate one-shot natural language spatial video grounding. We design an end-to-end model named IT-OS via information tree to address the challenges brought by limited labels.

- By leveraging the language query, several novel modules on the information tree, such as tree construction, branch search, and branch cropping, are proposed. Moreover, to strengthen the deep representation learning under limited supervision signals, we introduce several self-supervised tasks based on the information tree.

- We experiment with our IT-OS model on two benchmark datasets. Comparisons with the state-of-the-art and extensive model analysis jointly demonstrate the effectiveness of IT-OS.

## 2 Related works

**Natural Language Video Grounding.** Among numerous multimedia understanding applications (Zhang et al., 2020a,c, 2021d,c, 2020d; Kai et al., 2021; Zhang et al., 2020e), natural language video grounding has attracted the attention of more and more researchers recently. There are mainly three branches, temporal grounding[(Ross et al., 2018; Lu et al., 2019; Zhang et al., 2019; Lin et al., 2020a,b; Zhang et al., 2021a; Li et al., 2022; Gao et al., 2021; Yang et al., 2021)], spatio-temporal grounding[(Tang et al., 2021; Zhang et al., 2020f,g; Su et al., 2021)], and spatial grounding. We focus on the last one.

Deep neural network has convincingly demonstrated high capability in many domains (Wu et al., 2020, 2022; Guo et al., 2021; Li et al., 2020b,c,a), especially for video related tasks (Miao et al., 2021; Miao et al.; Xiao et al., 2020, 2021), like video

grounding. For example,(Li et al., 2017) use the neural network to detect language query related objects in the first frame and track the detected object in the whole video. Compared to it, (Yamaguchi et al., 2017) and (Vasudevan et al., 2018) go further. They extract all the object proposals through the pretrained detector, and choose the right proposal described in the text.

Supervised training for the natural language video object detection needs high labeling costs. To reduce it, some researchers pay attention to weakly-supervised learning fashion using multiple instances learning(MIL) method (Huang et al., 2018; Chen et al., 2019a; Shi et al., 2019; Chen et al., 2019b; Zhou et al., 2018; Wang et al., 2021a)transfers contextualized knowledge in cross-modal alignment to release the unstable training problem in MIL. Based on contrastive learning (Zhang et al., 2022), (Da et al., 2021) proposes an AsyNCE loss to disentangle false-positive frames in MIL, which allows for mitigating the uncertainty of from negative instance-sentence pairs. Weakly supervised false-positive identification based on contrastive learning has witnessed success in some other domains (Zhang et al., 2021b; Yao et al., 2022)

**One-shot Learning for Videos.** One-shot learning has been applied in some other video tasks. (Yang et al., 2018) proposes a meta-learning-based approach to perform one-shot action localization by capturing task-specific prior knowledge. (Wu et al., 2018) investigates the one-shot video person re-identification task by progressively improving the discriminative capability of CNN via stepwise learning. Different from these works, (Caelles et al., 2017) and (Meinhardt and Leal-Taixé, 2020) define the one-shot learning as only one frame being labeled per video. Specifically, (Caelles et al., 2017) use a fully convolutional neural network architecture to solve the one-shot video segmentation task. (Meinhardt and Leal-Taixé, 2020) decouple the detection task, and uses the modified Mask-RCNN to predict local segmentation masks. Following this setting, we investigate one-shot natural language spatial video grounding, and devise a novel information-tree based end-to-end framework for the task.

# 3 Method

## 3.1 Model Overview

**Problem Formulation.** Given a video $V = \{v^i\}_{i=1,2,...,I}$ and a natural language query $C$, spatial video grounding aims to localize the query-described object from all the objects $O^i = \{o^i_j\}_{j=1,2,...,J}$ for each frame. $I$ denotes the frame number of the video, and the $J$ is the object number in the video. In **one-shot** spatial video grounding, solely one frame $v^i$ in video $V$ is labeled with the region boxes of the target objects $O^i$.

**Pipeline of IT-OS.** As shown in Figure 2, there are mainly four steps involved in the end-to-end modeling of IT-OS:

- Firstly, we extract the features from the input video and the input caption. Specifically, for the video, we use ResNet-101(He et al., 2016) as the image encoder to extract the frame feature maps; for the language query, we employ a language model Roberta(Liu et al., 2019). Both the vision encoder and the language encoder are jointly optimized with the whole network.

- Secondly, we build the information tree to get the representation of the video. The information tree is built upon the frame feature maps, which are the leaf nodes. Leaf nodes will be further merged based on the relevance between node-node and node-query to have non-leaf and root nodes. Nodes on unnecessary branches will be deleted conditioned on the language query.

- Thirdly, we utilize the transformer encoder to reason on the remaining nodes and language features. Upon the transformer, we devise two self-supervised tasks, *i.e.*, masked feature modeling, and video-text matching, which enhances the representation learning under limited labels.

**Prediction and Training.** We follow the common prediction and training protocol of visual transformers used in other object detection models (Wang et al., 2021b). We input the embedding parameters $E_{de}$ and the multi-model features $F_{de}$ generated by the transformer encoder into the transformer decoder $D$. Then, the decoder $D$ outputs possible prediction region features for each frame. For each possible region, a possibility $P$ and a bounding box $B$ are generated.

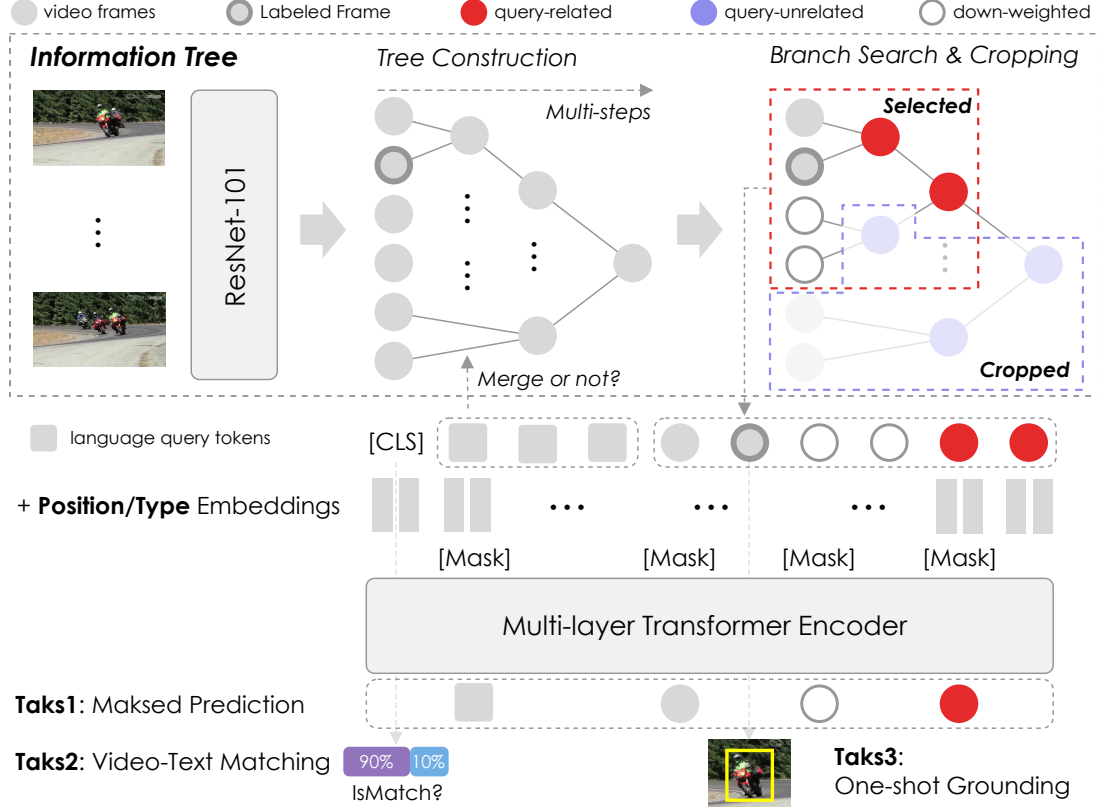$$P, B = D(F_{de}, E_{de}), \qquad (1)$$

Figure 2: The overall schema of the proposed end-to-end one-shot video grounding via information tree (IT-OS), which contains query-guided tree construction, query-based branch search & cropping, and a transformer encoder enhanced by self-supervised tasks.

We choose the box $B$ with the highest possibility value $P$ for each frame as the target box.

During the training process, we first calculate the possible prediction regions. Then, we match the possible regions with the target boxes, and choose the best match for each frame. Finally, use the match to train our IT-OS model.

### 3.2 Information Tree Module

In this section, we will elaborate the information tree modules in detail. We will illustrate how to construct the information tree, how to extract critical information from it and how to design the self-supervised learning based on the tree. To ease the illustration, we take the 6 frames as an example, and show the process in Figure 2.

#### 3.2.1 Tree Construction

Given the frame features generated by the CNN, we build the information tree by merging adjacent frame features in the specified order. Specifically, the frame features output by the image encoder are the leaf nodes $N = \{n^i\}_{i=1}^{2M}$. A sliding window of size 2 and step 2 is applied on these nodes and

nodes in the window are evaluated to be merged or not.

We calculate the *semantic relevance difference* between each node pair with the language query, and get the *visual relevance* between the nodes in each pair. For the visual relevance calculation, we max-pool the feature maps of the $i$ node pair to have the feature vector $f_v^{2i-1}$ and $f_v^{2i}$. And then, we compute the cosine similarity $r_{vv}^i$ between $f_v^{2i-1}$ and $f_v^{2i}$ to be the visual relevance. Next, we calculate the semantic relevance $r_{tv}^{2i-1}$ and $r_{tv}^{2i}$ between the text feature $f_t$ and the nodes of $i$ node pair:

$$r_{tv}^{2i-1} = \sigma((w_t * f_t) * (w_v * f_v^{2i-1})^T), \quad (2)$$

$$r_{tv}^{2i} = \sigma((w_t * f_t) * (w_v * f_v^{2i})^T), \quad (3)$$

where the $w_t$ and $w_v$ are learnable parameters, and $\sigma$ is the sigmoid activation function.

The semantic relevance difference $d_{tv}^i$ between the $i$th paired nodes is:

$$d_{tv}^i = |r_{tv}^{2i-1} - r_{tv}^{2i}| + \gamma * r_{vv}^i, \quad (4)$$

where the $\gamma$ is the hyperparameter.

With the relevant difference value, we rank the node pairs and pick out the top $\lambda$. The $\lambda$ is a hyperparameter, which can be set as a constant or a percentage. We merge the node pairs:

$$n^{new} = w_{mg} * (n^{2i-1} + n^{2i}) + b_{mg}, \quad (5)$$

where the $w_{mg}$ and $b_{mg}$ are trainable. Finally, The new node $n^{new}$ replace the old nodes $n^{2i-1}$ and $n^{2i}$ in the queue. Repeat the process until there is only one node in the queue. Saving all nodes in the process and the composite relationship between nodes generated in the merging process, we get the information tree.

### 3.2.2 Branch Search

We use a branch to denote a subtree. To filter critical local and global information, we perform branch search and selection. We firstly select branches that contain leaf nodes less than $\delta_{max}$ and more than $\delta_{min}$. $\delta_{max}$ and $\delta_{min}$ are hyperparameters. We calculate the *semantic relevance* of branches' root nodes and the language query based on Equation 2.

**Training.** During training, we directly select the branch that contains the labeled leaf node and the root node with the highest semantic relevance. This selection improves the training efficiency.

**Inference.** During inference, all frames should be processed. We conduct an iterative search with multiple search steps. For each step, we select the branch with the highest semantic relevance and remove the selected branch from the information tree. After the search, we have multiple selected branches and each branch will be forwarded to the following processes.

### 3.2.3 Branch Cropping

Note that not all the non-leaf nodes in the selected branches are closely related to the input caption. We remove non-leaf nodes that are with semantic relevance less than $\Delta$, which is a hyperparameter. Their descendant non-leaf nodes are also removed. To reserve enough frame nodes for training, we do not remove the descendant leaf nodes. Instead, we down-weight them with $\lambda = 0.5$. For other leaf nodes, $\lambda = 1$. The remaining leaf nodes and non-leaf nodes represent the critical local information and the global information, respectively. We multiply the feature of node $i$ and the node's semantic relevance $r_{tv}^i$:

$$f_{v_{new}}^i = f_v^i * r_{tv}^i * \lambda, \quad (6)$$

where $f_{v_{new}}^i$ is the feature vector input into the transformer. As such, Equation 6 considers both local relevance $r_{tv}$ and global relevance $\lambda$ with the language query.

### 3.2.4 Self-supervised Tasks

We leverage a transformer encoder for these extracted information and the language query. As shown in the Figure 2, we design two self-supervised tasks as: 1) predicting the masked text features, and masked local/global video information; 2) judging whether the text and the video match. For the transformer, the input tokens $F_{in}$ consist of the local information, the global information and the text features, which are three types of tokens. We further introduce 2-D position embedding for video tokens and type embedding for all tokens, which are added to the tokens' features.

Then, the features $F_{in}$ are input into the transformer encoder $E$. After encoding, the fusion features $F_{out}$ are output:

$$F_{out} = E(F_{in}). \quad (7)$$

We predict the original features for masked language tokens and masked video tokens (leaf/non-leaf nodes in the selected branch) using multilayer perceptrons.

$$\hat{f}_{in}^i = MLP_t(f_{out}^i), \quad \hat{f}_{in}^j = MLP_v(f_{out}^j), \quad (8)$$

where the $MLP_t$ and $MLP_v$ are the multilayer perceptrons for text and video features, respectively. We view masked token modeling as feature regression and adopt L2 distance as the loss function. In addition, there will be a mismatched language query at the rate of 50%. We propose to predict whether the video and language are matched, *i.e.*, whether the video contains the event described by the language query, based on the output representation of token [CLS]. When the video and the language are not matched, we will not train the model with the one-shot label.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets** We consider two video grounding benchmarks for evaluation: (1) *VidSTG (Zhang et al., 2020g)* is a large-scale benchmark dataset for video grounding, which is constructed based on VidOR (Shang et al., 2019) dataset. VidSTG contains $10,000$ videos and $99,943$ sentences with

| Method | Declarative Sentence Grounding | | | | Interrogative Sentence Grounding | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.4 | 0.5 | 0.6 | Avg | 0.4 | 0.5 | 0.6 | Avg |
| GroundeR | 24.56 | 18.22 | 13.73 | 18.85 | 25.28 | 18.87 | 14.39 | 19.52 |
| STPR | 25.68 | 20.07 | 14.64 | 19.89 | 27.09 | 21.04 | 16.00 | 21.38 |
| STGRN | 27.57 | 20.91 | 16.25 | 21.50 | 28.51 | 21.89 | 17.20 | 22.47 |
| VOGnet | 32.08 | 24.38 | 19.91 | 25.75 | 33.08 | 25.54 | 20.85 | 26.72 |
| OMRN | 34.43 | 27.57 | 21.91 | 27.96 | 35.69 | 28.74 | 23.03 | 29.14 |
| VOGnet* | 36.42 | 29.37 | 21.95 | 29.25 | 36.98 | 28.35 | 22.57 | 29.30 |
| OMRN* | 39.54 | 30.02 | 22.34 | 30.64 | 38.89 | 30.53 | 24.10 | 31.17 |
| **IT-OS** | **46.75** | **35.81** | **23.23** | **35.26** | **46.16** | **34.55** | **25.19** | **35.30** |

Table 1: Compared with baselines on VidSTVG. It is worth noting that all methods are trained using the one-shot learning. The $*$ represents the baselines use the MDETR as the object detector backbone, which is the same as the IT-OS.

| Method | 0.4 | 0.5 | 0.6 | Avg |
|---|---|---|---|---|
| GroundeR | 32.09 | 27.80 | 24.25 | 28.05 |
| STPR | 33.40 | 28.92 | 25.37 | 29.23 |
| STGRN | 35.45 | 30.41 | 26.31 | 30.72 |
| VOGnet | 38.81 | 32.65 | 26.87 | 32.78 |
| OMRN | 40.11 | 34.51 | 28.36 | 34.35 |
| VOGnet* | 41.23 | 35.82 | 29.48 | 35.51 |
| OMRN* | 45.52 | 37.69 | 30.41 | 37.87 |
| **IT-OS** | **51.87** | **42.91** | **33.58** | **42.79** |

Table 2: Compared with baselines on VID-sentence. All methods are trained using one-shot learning. The $*$ represents the MDETR is applied to these baselines as the object detector backbone.

$55, 135$ interrogative sentences and $44, 808$ declarative sentences. These sentences describe 79 types of objects appearing in the videos. We follow the official dataset split of (Zhang et al., 2020g). (2) *VID-sentence (Chen et al., 2019b)* is another widely used video grounding benchmark constructed based on the VID (Russakovsky et al., 2015) dataset. There are 30 categories and $7, 654$ video clips in this dataset. We report the results of all methods on the validation set for the VID-sentence dataset. We obtain similar observations and conclusions on the test set.

**Implementation Detail** For video preprocessing, we random resize the frames, and set the max size is $640 * 640$. The other data augmentation methods, such as random horizontal flip and random size cropping are used at the same time. During training, the learning rate is by default 0.00005, and decays by a factor of 10 for every 35 epochs. The batch size is 1 and the maximum training epoch is 100. We implement IT-OS in Pytorch and train

it on a Linux server. For model hyperparameters, we set $\lambda = 60\%$, and $\Delta = 0.7$. Most of the natural language spatial video grounding models use the pretrained detection model as the backbone. Thus, like them, we choose the official pretrained MDETR (Kamath et al., 2021) as the parameter basis for target detection of our IT-OS.

**Evaluation Metrics** We follow the evaluation protocol of (Chen et al., 2019b). Specifically, we compute the Intersection over Union (IoU) metric for the predicted spatial bounding box and the ground-truth per frame. The prediction for a video is considered as "accurate" if the average IoU of all frames exceeds a threshold $\alpha$. The $\alpha$ is set to 0.4, 0.5, and 0.6 during testing.

**Baselines** Since existing video grounding methods are not directly applicable to the one-shot setting, we extend several state-of-the-arts as the baselines. Specifically, to have a comprehensive comparison, we consider 1)fully supervised models, including **VOGnet** (Sadhu et al., 2020), **OMRN** (Zhang et al., 2020f) and **STGRN** (Zhang et al., 2020g); and 2) other widely known methods, including video person grounding **STPR** (Yamaguchi et al., 2017), and visual grounding method, **GroundeR** (Rohrbach et al., 2016).

### 4.2 Performance Comparison

The experimental results for one-shot video grounding on VidSTVG and VID-sentence datasets are shown in Table 1 and 2, respectively. According to the results, we have the following observations:

- Not surprisingly, although extended to the video grounding setting, baselines that belong to other domains, including video person grounding STPR and visual grounding GroundeR, achieve

inferior results on video grounding benchmarks. They lack domain-specific knowledge and might fail to effectively model the spatial-temporal relationships of videos and language queries.

- IT-OS consistently achieves the best performance on two benchmarks and multiple experimental settings with a large margin improvement. Remarkably, IT-OS boosts the performance (Avg) of the previous state-of-the-art OMRN from nearly $28.0/29.1/34.4$ to $35.3/35.3/42.8$ on VidSTVG and VID-sentence, respectively. It demonstrates the superiority of IT-OS on one-shot video grounding.

- The baselines are implemented with the backbones used in their original papers, which are different from ours. To further disentangle the sources of performance improvement, we re-implement the best-performing baselines (VOGnet*, and OMRN*) with the same object detection backbone, MDETR, as IT-OS. Although there is performance improvement with the new backbone, the best-performing baseline OMRN*, still underperforms IT-OS by over $4$ points for the average accuracy on all datasets. It further reveals the effectiveness of our novel model designs eliminating interference with different pre-training parameters. We attribute the improvement to the end-to-end modeling, where different modules can simultaneously benefit from each other. In addition, the proposed information tree alleviates the negative effects of irrelevant frames, and effectively models the interactions between the video global/local information and the language query. Several self-supervised learning tasks based on the information tree enhance the representation learning under limited one-shot labels.

### 4.3 Comparison with Fully Supervised Methods

We are interested in 1) how different baselines perform under fully supervised settings; 2) how one-shot IT-OS perform compared to these baselines. Towards this end, we train multiple baselines and IT-OS with all labels on the VID-sentence dataset. The experiment results are shown in Table 3. From the table, we have the following findings:

- Remarkably, the performance gap between one-shot IT-OS and the fully supervised OMRN is

| Method | 0.4 | 0.5 | 0.6 | Avg |
|---|---|---|---|---|
| GroundeR | 42.72 | 33.77 | 27.05 | 34.51 |
| STPR | 47.95 | 36.19 | 30.41 | 38.18 |
| STGRN | 49.25 | 44.03 | 34.89 | 42.72 |
| VOGnet | 53.17 | 43.47 | 33.77 | 43.47 |
| OMRN | 55.22 | 46.64 | 37.50 | 46.45 |
| IT-OS (OS) | 51.87 | 42.91 | 33.58 | 42.79 |

Table 3: Compared with the baselines on VID-sentence. The baselines are trained using fully supervised learning. The OS represents the IT-OS trained under the one-shot settings.

less than $4\%$. Such a minor gap demonstrates the effectiveness of IT-OS on learning with limited annotations. This is significant and practical merit since we are more likely to have a limited annotation budget in real-world applications.

- Surprisingly, one-shot IT-OS can still outperform some weak baselines such as GroundeR and STPR. These results reveal the necessity of end-to-end modeling for video grounding.

### 4.4 Ablation Study

We are interested in how different building blocks contribute to the effectiveness of IT-OS. To this end, we surgically remove several components from IT-OS and construct different architectures. The investigated components include information tree ($\Gamma_{tree}$), the branch cropping ($\Gamma_{crop}$), and the self-supervised training ($\Gamma_{self}$). It is worth noting that the other components cannot be deleted independently except the branch cropping. Thus, we don't conduct an ablation study for them. Results on VidSTG and VID-sentence datasets are shown in Table 4 and Table 5, respectively. There are several observations:

- Overall, removing any component incurs a performance drop, demonstrating the necessity and effectiveness of the information tree, branch search & cropping, and self-supervised training.

- Stacking multiple components outperform the architecture with a single component. This result reveals that the proposed components can benefit from each other in end-to-end training and jointly boost one-shot video grounding.

### 4.5 Case Study

We conduct a case study to visually reveal the ability of the IT-OS in detail. Specifically, we random

| $\Gamma_{self}$ | $\Gamma_{tree}$ | $\Gamma_{crop}$ | Declarative Sentence Grounding | | | | Interrogative Sentence Grounding | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0.4 | 0.5 | 0.6 | Avg | 0.4 | 0.5 | 0.6 | Avg |
| | | | 39.00 | 30.52 | 17.61 | 29.05 | 38.78 | 28.75 | 19.67 | 29.07 |
| ✓ | | | 40.52 | 32.32 | 18.83 | 30.56 | 40.82 | 31.44 | 20.66 | 30.97 |
| | ✓ | | 42.34 | 32.65 | 20.35 | 31.78 | 42.26 | 32.02 | 21.89 | 32.06 |
| ✓ | ✓ | | 44.16 | 33.38 | 21.11 | 32.89 | 44.55 | 33.78 | 23.19 | 33.84 |
| | ✓ | ✓ | 44.77 | 34.62 | 22.93 | 34.11 | 44.30 | 33.23 | 24.17 | 33.90 |
| ✓ | ✓ | ✓ | **46.75** | **35.81** | **23.23** | **35.26** | **46.16** | **34.55** | **25.19** | **35.30** |

Table 4: Ablation study on VidSTG dataset.

| $\Gamma_{self}$ | $\Gamma_{tree}$ | $\Gamma_{crop}$ | 0.4 | 0.5 | 0.6 | Avg |
|---|---|---|---|---|---|---|
| | | | 44.40 | 35.07 | 27.24 | 35.57 |
| ✓ | | | 46.64 | 36.38 | 28.54 | 37.19 |
| | ✓ | | 47.95 | 38.99 | 29.85 | 38.93 |
| ✓ | ✓ | | 49.44 | 40.30 | 31.16 | 40.30 |
| | ✓ | ✓ | 50.19 | 40.49 | 32.46 | 41.04 |
| ✓ | ✓ | ✓ | **51.87** | **42.91** | **33.58** | **42.79** |

Table 5: Ablation study on VID-sentence dataset.

sample 3 videos from the datasets, and sample 6 frames from each video to visualize.

We compare our IT-OS model with the baseline method, OMRN, and the fundamental ablation model of the IT-OS, which is removed from the self-supervised module and the information tree. As shown in Figure 3, we have the following key findings: (1) The IT-OS detects the more accurate one from all objects of the video than the best performing previous method. It demonstrates the better representation extraction and analysis capabilities of our model. (2) Even if the target object is selected correctly, the IT-OS localizes a more precise spatial area compared with the previous two stages method. The results reflect the end-to-end model, IT-OS, has more accurate domain knowledge through training the whole model on the target dataset. (3) After adding the information tree and the self-supervised module, the IT-OS outputs more precise bounding boxes. It reveals that combining the two modules introduce stronger supervision signals for model training so that the model has stronger detection ability.

## 5 Conclusion

In this paper, we introduce the one-shot learning into the natural language spatial video grounding task to reduce the labeling cost. To achieve the goal, the main point is to make full use of only one frame label for each video. The invalid frames
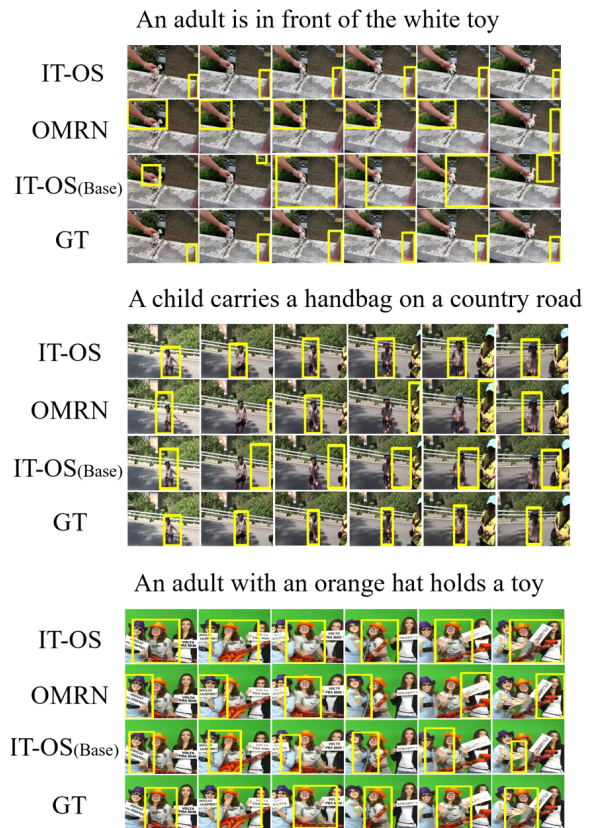


An adult is in front of the white toy

IT-OS

OMRN

IT-OS(Base)

GT

A child carries a handbag on a country road

IT-OS

OMRN

IT-OS(Base)

GT

An adult with an orange hat holds a toy

IT-OS

OMRN

IT-OS(Base)

GT

Figure 3: Examples of the detection result visualization. The IT-OS(Base) represents the IT-OS model without the self supervised module and the informaiton tree. The GT represents the target labels.

unrelated to the input text and target objects bring confounding to the one-shot training process. We design an end-to-end model (IT-OS) via the information tree to avoid it. Specifically, the information tree module merges frames with similar semantics into one node. Then, by searching the tree and cropping the invalid nodes, we can get the complete and valid semantic unit of the video. Finally, two self-supervised tasks are used to make up the insufficient supervision.

## Acknowledgements

## References

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2017. One-shot video object segmentation. In *CVPR*.

Lei Chen, Mengyao Zhai, Jiawei He, and Greg Mori. 2019a. Object grounding via iterative context reasoning. In *ICCV Workshops*.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. 2019b. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv*.

Cheng Da, Yanhao Zhang, Yun Zheng, Pan Pan, Yinghui Xu, and Chunhong Pan. 2021. Asynce: Disentangling false-positives for weakly-supervised video grounding. In *MM*.

Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. 2021. Relation-aware video reading comprehension for temporal language grounding. In *EMNLP*.

Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. 2021. Semi-supervised active learning for semi-supervised models: exploit adversarial examples with graph-based virtual labels. In *ICCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *ACL*.

De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In *CVPR*.

Shen Kai, Lingfei Wu, Siliang Tang, Yueting Zhuang, zhen he, Zhuoye Ding, Yun Xiao, and Bo Long. 2021. Learning to generate visual questions with noisy supervision. In *NIPS*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*.

Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. 2021. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *ICCV*.

Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. 2020a. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *CVPR*.

Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*.

Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu. 2020b. Multi-task attribute-fusion model for fine-grained image recognition. In *Optoelectronic Imaging and Multimedia Technology VII*.

Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. 2020c. Ib-m: A flexible framework to align an interpretable model and a black-box model. In *BIBM*.

Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. 2017. Tracking by natural language specification. In *CVPR*.

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

Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. 2020b. Moment retrieval via cross-modal interaction networks with query reconstruction. *TIP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

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*.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tim Meinhardt and Laura Leal-Taixé. 2020. Make one-shot video object segmentation efficient again. *arXiv*.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. 2021. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*.

Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.

Candace Ross, Andrei Barbu, Yevgeni Berzak, Battushig Myanganbayar, and Boris Katz. 2018. Grounding language acquisition by training semantic parsers using captioned videos. In *EMNLP*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video object grounding using semantic roles in language description. In *CVPR*.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *ICMR*.

Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. 2019. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*.

Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *CVPR*.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *TCSVT*.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. In *CVPR*.

Wei Wang, Junyu Gao, and Changsheng Xu. 2021a. Weakly-supervised video object grounding via stable context learning. In *MM*.

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021b. End-to-end video instance segmentation with transformers. In *CVPR*.

Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yue Ting Zhuang, and Fei Wu. 2022. Learning decomposed representations for treatment effect estimation. *TKDE*.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *EMNLP*.

Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *ECCV*.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2021. Video as conditional graph hierarchy for multi-granular question answering. *arXiv*.

Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Spatio-temporal person retrieval via natural language queries. In *ICCV*.

Hongtao Yang, Xuming He, and Fatih Porikli. 2018. One-shot action localization by learning sequence matching network. In *CVPR*.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *SIGIR*.

Dong Yao, Zhou Zhao, Shengyu Zhang, Jieming Zhu, Zhu Yudong, Rui Zhang, and Xiuqiang He. 2022. Contrastive learning adopting positive-negative frame mask for music representation. In *WWW*.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021a. Parallel attention network with sequence matching for video grounding. *arXiv*.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020a. Devlbert: Learning deconfounded visio-linguistic representations. In *MM*.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020b. Poet: Product-oriented video captioner for e-commerce. In *MM*.

Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020c. Comprehensive information integration modeling framework for video titling. In *KDD*.

Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2022. Re-contrast, re-attend, re-construct for multi-interest recommendation. In *WWW*.

Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021b. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *SIGIR*.

Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. 2021c. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. *arXiv*.

Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021d. Consensus graph representation learning for better grounded image captioning. In *AAAI*.

Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. 2020d. Relational graph learning for grounded video description generation. In *MM*.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020e. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NIPS*.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Nicholas Jing Yuan. 2020f. Object-aware multi-branch relation networks for spatio-temporal video grounding. *arXiv*.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020g. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. 2019. Grounded video description. In *CVPR*.

Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv*.