The Supplementary Material —— Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences

1. Dataset Details

VidOR contains 7,000, 835 and 2,165 videos for training, validation and testing, respectively. Since box annotations of testing videos are unavailable yet, we omit testing videos, split 10% training videos as our validation data and regard original validation videos as the testing data. Considering a video may contain multiple same triplets that have different temporal and bounding box annotations, we cut these videos into several short videos, where each short video contains a triplet that covers a segment of the short video. We then delete unsuitable video-triplet pairs based on three rules: (1) the video length is less than 3 seconds; (2) the temporal duration of the triplet is less than 0.5 seconds; (3) the triplet duration is less than 2% of the video. Next, because too many triplets are related to spatial relations like "in_front_of" and "next_to", we delete 90% spatial triplets to keep the types of relations balanced.

For each video-triplet pair, we choose the $subject$ or $object$ as the queried object, and then describe its appearance, relationships with other objects and visual environments. We discard video-triplet pairs that are too hard to give a precise description. And a video-triplet pair may correspond to multiple sentences. After annotation, there are 6,924 videos (5,563, 618 and 743 for training, validation and testing sets) and 99,943 sentences for 44,808 video-triplet pairs. we show some typical samples in Figure 1 with declarative and interrogative sentences. We can find that the objects may exist in a very small segment of the video and the sentences may only describe a short-term state of the queried object.

Next, we show the distribution of different types of queried objects as Figure 2. The original VidOR contains 80 types of objects, including 3 types of persons, 28 types of animals and 49 types of other objects. After data cleaning and annotating, sentences in VidSTG describes 79 types of objects by the declarative or interrogative ways, including 3 types of persons, 27 types of animals and 49 types of other objects. A rare type (i.e., stingray) is not contained in VidSTG. From Figure 2, we can find the sentences of person types take up the largest proportion and sentence numbers of other categories are relatively uniform.

Moreover, we compare VidSTG with existing video grounding datasets in Table 1. Previous temporal sentence grounding datasets like DiDeMo [4], Charades-STA [3], TACoS [7] and ActivityCation [6] only provide the temporal annotations for each sentence and lack the spatio-temporal bounding boxes. As for existing video grounding datasets, Persen-sentence [10] is originally used for spatio-temporal person retrieval among trimmed videos and only contains one type of objects (i.e. people), which is too simple for the STVG task. And VID-sentence dataset [2] contains 30 categories but also offer the annotations on trimmed videos. Different from them, our VidSTG simultaneously offers temporal clip and spatio-temporal tube annotations, contains more sentence descriptions, has a richer variety of objects, and further supports multi-form sentences.

2. Baseline Details

Since no existing strategy can be directly applies to STVG, we combine the existing visual grounding method **GroundeR** [8] and spatio-temporal video grounding approaches **STPR** [10] and **WSSTG** [2] with the temporal sentence localization methods **TALL** [3] and **L-Net** [1] as the baselines. The TALL and L-Net first provide the temporal clip of the target tube and the extended GroundeR, STPR and WSSTG then retrieve the spatio-temporal tubes of objects.

We first introduce the TALL and L-Net approaches. The TALL applies a sliding window framework that first samples abundant candidate clips and then ranks them by estimating the clip-sentence scores. During estimating, TALL incorporates the context features for the current clip to further improve the localization accuracy. And L-Net develops the evolving frame-by-word interactions for video and query contents, and dynamically aggregates the matching evidence to localize the temporal boundaries of clips according to the textual query.

Next, we illustrate the extended grounding methods GroundeR, STPR and WSSTG based on the retrieved clip. The GroundeR is a frame-level approach, which originally grounds natural language in a still image. We apply it for each frame of the clip to obtain the object region and generate a tube by directly connecting these regions. The drawback of this method is the lack of temporal context

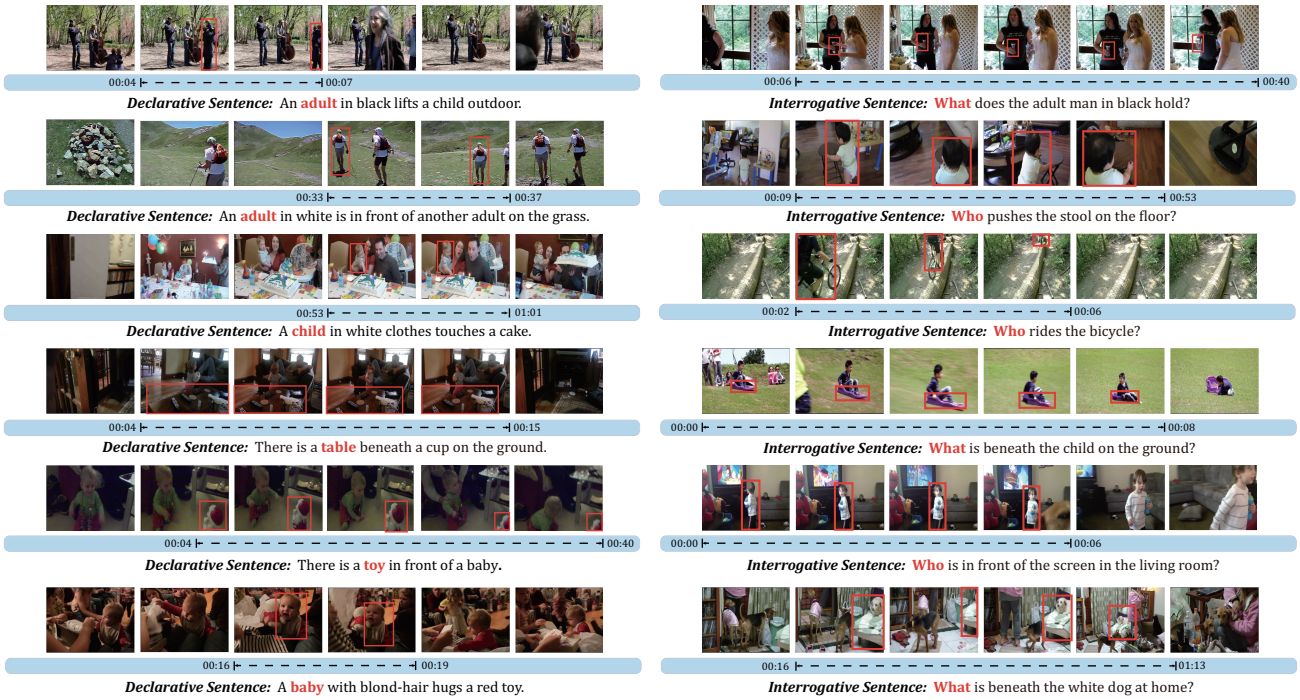

Figure 1. Annotation Samples with Declarative or Interrogative Sentence Descriptions.

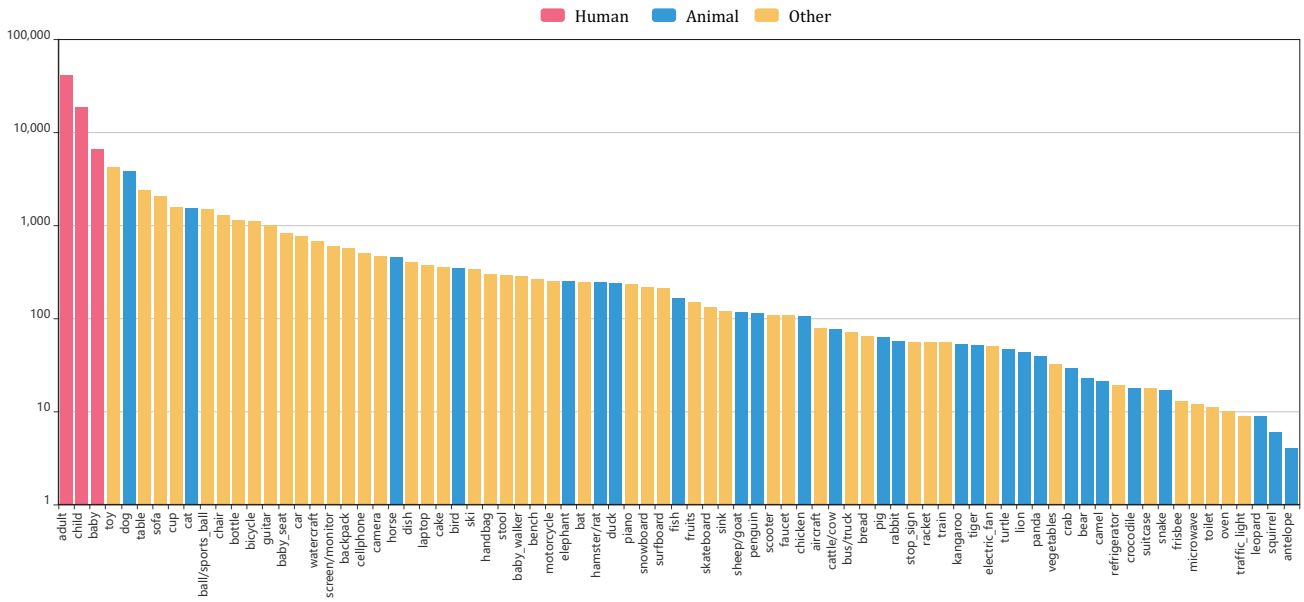

Figure 2. The Distribution of Different Types of Queried Objects in the Entire VidSTG Dataset.

modeling of regions. Different from it, original STPR and WSSTG are both tube-level methods and adopt the tube pre-generation framework. This framework first extracts a set of spatio-temporal tubes from trimmed clips and then identifies the target tube. The original STPR [10] only grounds persons from multiple videos, we extend it to multi-type object grounding in a single clip. Specifically, we use the pre-trained Faster R-CNN to detect multi-type object regions to generate the candidate tubes rather than only generate person candidate tubes. And during training, we retrieve the correct tube from a video rather multiple videos, where we do not change the loss func-

Table 1. Dataset Comparison.

Dataset	#Video	#Sentence	#Type	Domain	Temporal Ann.	Box Ann.	Multi-Form Sent.
TACoS	127	18,818	-	Cooking	✓		
DiDeMo	10,464	40,543	-	Open	✓		
Charades-STA	6,670	16,128	-	Activity	✓		
ActivityCaption	20,000	37,421	-	Open	✓		
Person-sentence	5,293	30,365	1	Person		✓	
VID-sentence	4,318	7,654	30	Open		✓	
VidSTG	6,924	99,943	79	Open	✓	✓	✓

Table 2. Ablation Results of Directed GCN.

Method	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
w/o. Explicit Subgraph	46.70%	18.07%	22.23%	13.12%
Undirected GCN	46.81%	18.13%	22.28%	13.06%
Undirected GAT	47.08%	18.21%	22.35%	13.20%
Directed GCN (our)	**47.64%**	**18.96%**	**23.19%**	**13.62%**

Table 3. Ablation Results of Query Modeling.

Method	Query Modeling	m_tIoU	m_vIoU	vIoU@0.3
WSSTG+L-Net	GRU	40.27%	13.85%	17.66%
	GRU+Object Rec.+Attention	41.22%	14.32%	20.08%
STGRN	GRU	46.93%	18.42%	22.41%
	GRU+Object Rec.+Attention	**47.64%**	**18.96%**	**23.19%**

tion of STPR. The original WSSTG [2] employs a weakly-supervised setting, we extend it to the fully-supervised form. Concretely, we discard the original ranking and diversity losses and employs a classics triplet loss [11] on the matching scores of the candidate tubes and sentence. The STPR and WSSTG both have the drawbacks of the tube pre-generation framework: (1) they are hard to pre-generate high-quality tubes without textual clues; (2) they only consider single tube modeling and ignore the rela1ionships between objects. Finally, we obtain 6 combined baselines **GroundeR+TALL**, **STPR+TALL**, **WSSTG+TALL**, **GroundeR+L-Net**, **STPR+L-Net** and **WSSTG+L-Net**. We also provide the temporal ground truth clip to form 3 baselines **GroundeR+Tem.Gt**, **STPR+Tem.Gt** and **WSSTG+Tem.Gt**.

During training, we first train the TALL and L-Net based on the sentence-clip matching data and train GroundeR, STPR and WSSTG within the ground truth clip. But while inference, we first use TALL and L-Net to determine the clip boundaries and then employ GroundeR, STPR and WSSTG to localize the final tubes. To guarantee the fair comparison, GroundeR, STPR and WSSTG are built on the same TALL or L-Net models, and we apply the Adam optimizer to train all baselines.

3. More Ablation Study

3.1. Directed GCN

To confirm the effect of the directed explicit GCN, we replace it with the original undirected GCN [5] and GAT [9]. In Table 2, our directed GCN has a better performance and the results of undirected GCN and GAT are close to the model without explicit subgraph modeling. The reason is that the undirected GCN and GAT have a similar ability with the implicit GCN and may lead to redundancy modeling.

3.2. Query Modeling

Our input setting is consistent with previous grounding works but we adopt a different strategy for query modeling in STGRN. Previous works model the sentence by RNN as a whole query vector. Different from them, we use the NLTK library to recognize the first noun or interrogative word "who/what" in the sentence, corresponding to the query object. We then select its feature $\mathbf{s}^e$ from RNN outputs and adopt context attention to learn the object-aware query vector $\mathbf{s}^q$. We conduct an ablation study for query modeling in Table 3, where we also apply the **GRU+Object Rec.+Attention** to WSSTG+L-Net. Concretely, the object-aware vector $\mathbf{s}^q$ replaces the original sentence vector in final localization for L-Net and is added into visually guided sentence features for WSSTG.

References

[1] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019.

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

[3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285. IEEE, 2017.

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

[5] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

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

[7] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association of Computational Linguistics*, 1:25–36, 2013.

[8] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, pages 817–834. Springer, 2016.

[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2017.

[10] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *ICCV*, pages 1453–1462, 2017.

[11] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *CVPR*, pages 4145–4154, 2019.