Video-Text Prompting for Weakly Supervised Spatio-Temporal Video Grounding

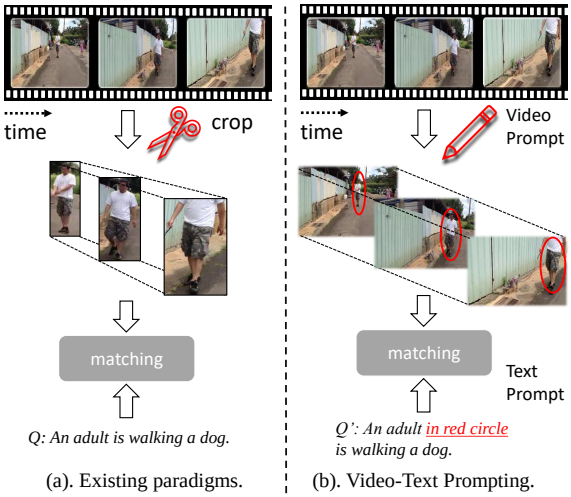
Heng Zhao[*1,2,4], **Yinjie Zhao**[*1,2,3], **Bihan Wen**[3], **Yew-Soon Ong**[1,2,4], **Joey Tianyi Zhou**[1,2]
[1] CFAR, Agency for Science, Technology and Research(A*STAR), Singapore
[2] IHPC, Agency for Science, Technology and Research(A*STAR), Singapore
[3] School of EEE, Nanyang Technological University, Singapore
[4] CCDS, Nanyang Technological University, Singapore

Abstract

Weakly-supervised Spatio-Temporal Video Grounding(STVG) aims to localize target object tube given a text query, without densely annotated training data. Existing methods extract each candidate tube feature independently by cropping objects from video frame feature, discarding all contextual information such as position change and inter-entity relationship. In this paper, we propose Video-Text Prompting(VTP) to construct candidate feature. Instead of cropping tube region from feature map, we draw visual markers(e.g. red circle) over objects tubes as video prompts; corresponding text prompt(e.g. *in red circle*) is also inserted after the subject word of query text to highlight its presence. Nevertheless, each candidate feature may look similar without cropping. To address this, we further propose Contrastive VTP(CVTP) by introducing negative contrastive samples whose candidate object is erased instead of being highlighted; by comparing the difference between VTP candidate and the contrastive sample, the gap of matching score between correct candidate and the rest is enlarged. Extensive experiments and ablations are conducted on several STVG datasets and our results surpass existing weakly-supervised methods by a great margin, demonstrating the effectiveness of our proposed methods.

1 Introduction

The task of STVG is of high importance to real world applications such as general artificial intelligence for video understanding, information retrieval for surveillance systems and human-machine interaction, etc. To relieve the reliance on heavily annotated training data, weakly-supervised STVG deserves more research attention. Regrettably, training under this setting is an extremely challenging because only the pairing information between video clips and their corresponding query texts is available during training and no bounding box nor temporal annotation could be used.



(a). Existing paradigms. (b). Video-Text Prompting.

Figure 1: Comparison between existing paradigms (Chen et al., 2019b; Li et al., 2023) and the proposed Video-Text Prompting(VTP). Given pre-extracted candidate tube boxes, (a) existing methods obtain candidate feature by cropping from frame feature, resulting in contextual information loss; (b) our proposed VTP constructs context-preserving candidates by adding visual and textual prompts to the input.

It is common for weakly-supervised methods to reformulate the grounding problem as a retrieval task where candidate tube boxes are obtained with pre-trained object detectors and trackers then later ranked based on the query to find the best pick. Existing methods reason with the entity's tube feature cropped from the global frame feature given its boxes obtained in pre-processing step, discarding all contextual information such as the entity's relationship with other entities, its moving trajectory and so on(illustrated in Fig. 1 (a)). We note that this is an inherent difficulty for existing methods as their feature extraction is usually done together with the pre-processing step.

To overcome this weakness, we propose Video-

[*]The authors contributed equally to this work.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19494–19505
November 12-16, 2024

Text Prompting(VTP). In detail, by transforming the tube boxes obtained in pre-processing into video prompt(e.g. red circles) that are drawn as markers on the input video frames, as shown in Fig. 1 (b); we are able to preserve all useful context information for reasoning. Meanwhile we prompt the query sentence correspondingly by inserting text prompt(e.g. *in red circle*) after the subject word. We refer to candidate instances created in this way as Video Prompted(VP) candidates. Notably, this will create a visual discrepancy with the prompted query if the visual prompt lands on an incorrect candidate thus lowering its matching potential. Nevertheless, this discrepancy can be subtle as it may contradict with the prompted query partially. For example, the VP candidate 2 in Fig. 2 is an interference as it partially matches the prompted query: *An adult in red circle?* (Yes), *He is walking a dog?* (No).

To address this issue, we further propose Contrastive VTP(CVTP) where we construct a Contrastive Video Prompted(CVP) counterpart for each VP candidate by erasing its presence rather than highlighting it. Intuitively, the matching score of the CVP counterpart will be decimated in case the correct candidate is erased and meanwhile it is high for those incorrect candidate who is irrelevant to the query text. Thus by calculating the matching score difference between VP candidates and its CVP counterparts, we suppress the interference candidates and enlarge the gap between the correct candidate and the incorrect ones, as shown in Fig. 2.

Another challenge in weakly-supervised STVG is the temporal reasoning required to refine the selected candidate tube along the time axis. We address this by implementing a light-weight encoder-decoder transformer where the self-attention layers in encoder models temporal interaction between the prompted video frames and the cross-attention layers in decoder models the multi-modal reasoning between visual and linguistic feature. We conduct extensive experiments on two commonly used STVG datasets and our results surpasses existing weakly-supervised counterparts by a notable margin, which demonstrates its effectiveness.

We list our contributions as follows:

- To our knowledge, we are the first to explore video-text prompting, creating artificial local emphasis without losing global contextual information; which is especially beneficial for video related tasks with complex multi-entity interaction.
- We propose a novel Contrastive Video-Text Prompting method for weakly-supervised STVG to create contrast between videos with highlighted and erased candidate information, enlarging the advantage of the correct candidate in ranking.
- Our method achieves SOTA performance by a margin on widely used datasets and certain results even outperform some of the supervised methods.

2 Related Works

Visual Prompting Originated for NLP community, prompting methodology can be generalized as adding fixed or trainable parameters to the raw input. Most of the early prompting inspired works for visual-related tasks (Radford et al., 2021; Zhou et al., 2022; Ju et al., 2022) only prompt text or class labels, there is also works explore visual prompts by adding learnable pixels or tokens to raw image input (Bahng et al., 2022; Wu et al., 2022; Jia et al., 2022). The form of visual prompts become diverse in recent years, such as bounding boxes (Yao et al., 2021), masks (Li et al., 2024) or even mouse clicks (Kirillov et al., 2023). Although the different prompt types, these work treat visual prompts as a visual prior or anchor to better understand the prompted region. In contrast, we use visual prompts as an local emphasis to contrast with other entities. The most similar work to our setup is (Shtedritski et al., 2023), which also uses red circle as visual prompts to highlight the region for fine-grained local perception. However, our proposed CVTP not just employs visual prompts, it further leverages corresponding prompted texts and CVP counterpart to empower the contrast between candidates.

Weakly Supervised Video Grounding Fully-supervised STVG methods (Zhang et al., 2020b; Su et al., 2021; Yang et al., 2022; Jin et al., 2022; Lin et al., 2023) hold SOTA performance with a large margin compared with weakly-supervised counterparts; Nevertheless, the requirements of frame-level bounding box annotation and temporal boundary with second-level precision is impractical when the model needs to be trained for a new application with different data distributions. However, research

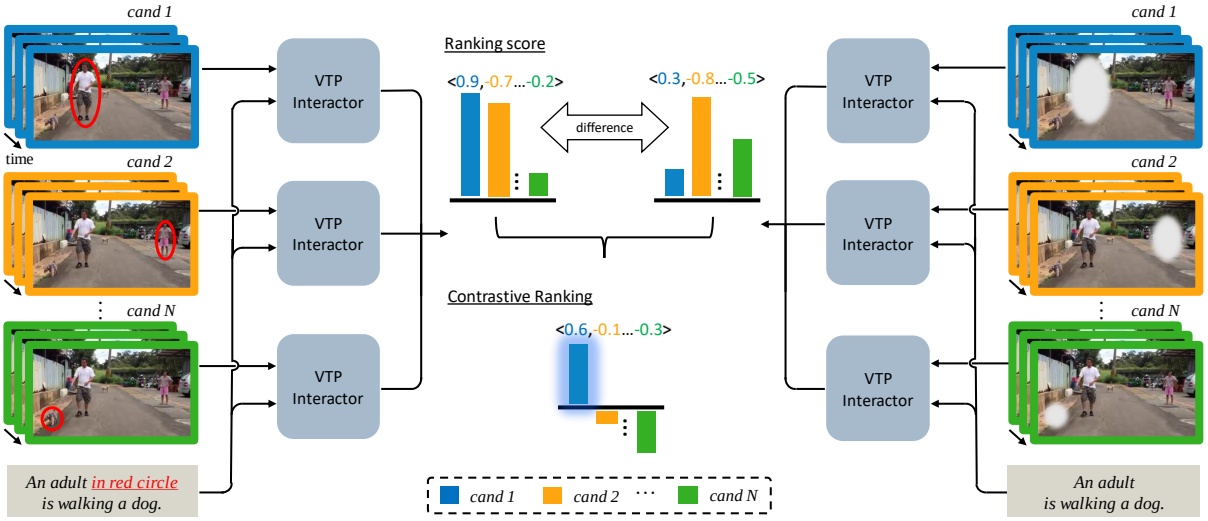


Figure 2: Contrastive Video-Text Prompting(CVTP) framework. VTP constructs candidates by video-text prompting and pick the highest ranked one while their CVP counterparts can also be ranked to pick the lowest ranked one. Contrastively, The CVTP framework is able to magnify the ranking score difference between correct and incorrect candidates.

for STVG under weakly-supervised setting is not receiving enough attention. (Shi et al., 2019) calculates video-text similarity score by averaging frame-level region-query scores without considering temporal cues. (Chen et al., 2020) breaks down query into object and activity to model a finer matching behavior between region embeddings. (Chen et al., 2019b) leverages LSTM (Hochreiter and Schmidhuber, 1997) to model temporal interaction of an extracted tube candidate. (Li et al., 2023) builds a language decomposition tree with the query to perform hierachical video-text alignment. However, all existing methods uses local feature for each candidate by cropping the entity region out from the whole frame. In this process, all contextual information is lost such as entity relative position and their possible directional interactions. In contrast, our VTP method keep the context intact.

3 Methodology

3.1 Preliminary

Spatial-temporal Video Grounding (STVG) STVG aims to localize a spatial-temporal tube $\mathcal{P} = \{\boldsymbol{b}_t\}_{t=t_s}^{t_e}$ in an untrimmed video $\mathcal{V} = \{\boldsymbol{v}_t\}_{t=1}^{T}$ with a given query sentence $\mathcal{S} = \{s_m\}_{m=1}^{L}$, where $\boldsymbol{b}_t$ is a bounding box for video frame t spanning from starting frame t_s to ending frame t_e, $\boldsymbol{v}_t$ is the frame-level input for video $\mathcal{V}$, and s_m is the token-level notation query sentence $\mathcal{S}$, respectively.

Weakly Supervised STVG In this setting, grounding is commonly handled as a retrieval task where only video-sentence pair-wise correspondence $(\mathcal{V}, \mathcal{S})$ is available. Specifically in pre-processing step, a set of candidate tube boxes $\{\mathcal{P}_i\}_{i=1}^{N_p}$ are extracted and their feature are cropped at the same time $\{\hat{\mathcal{P}}_i\}_{i=1}^{N_p}$. For one video clip $\mathcal{V}$ there is only one out of N_p tubes that is considered as the correct candidate. The core of the retrieval objective is to train a scoring or similarity function $\phi(\cdot)$ to rank the candidates $\{\mathcal{P}_i\}$ based on the query $\mathcal{S}$ whose feature is $\hat{\mathcal{S}}$. The video-text similarity score $s(\cdot)$ and best matching candidate index i^* is expressed as follows:

$$s(\mathcal{V}, \mathcal{S}) = \max_i \phi(\hat{\mathcal{P}}_i, \hat{\mathcal{S}}) \quad (1)$$

$$i^* = \arg\max_i \phi(\hat{\mathcal{P}}_i, \hat{\mathcal{S}}) \quad (2)$$

the ranking function $\phi(\cdot)$ is usually learnt via a contrastive loss between the positive sample pair $(\mathcal{V}, \mathcal{S})$ and the negative sample pairs $(\mathcal{V}', \mathcal{S})$ and $(\mathcal{V}, \mathcal{S}')$, under a Multiple-Instance Learning(MIL) paradigm (Karpathy and Fei-Fei, 2015).

3.2 Our approach

In general, the ranking function $\phi(\cdot)$ measures the similarity between the tube candidate's visual feature and the query sentence's feature in the latent space. Existing methods usually extract such feature via pre-trained uni-modality models such as Faster RCNN (Ren et al., 2015) and BERT (Devlin

et al., 2019). However, there are two shortcomings. For one, the visual feature space and linguistic feature space is not aligned because the object-detector and the language embedding model are trained separately with uni-modality data only. And most importantly, extracting candidate's feature with an object-detector usually involves a cropping operation(e.g. RoI Pooling(Girshick, 2015) and RoI Align(He et al., 2017)) on the frame-level feature map: $\mathcal{P}_i = \mathrm{Crop}(\mathcal{V}, \mathcal{P}_i) = \{\mathrm{Crop}(\boldsymbol{v}_t, \boldsymbol{b}_t^i)\}_{t=t_s}^{t_e}$. As such, the visual feature for one candidate tube can only carry its own information, neglecting all other contextual information that could be important for reasoning.

To address above weaknesses, we propose Video-Text Prompting(VTP) as a baseline which is further extended by our Contrastive Video-Text Prompting(CVTP) framework.

Video-Text Prompting Pre-trained large Vision-Language foundation Models(VLMs) such as CLIP(Radford et al., 2021) have proven to be strong multi-modal feature extractors. The feature gap issue can be addressed by employing VLMs as backbones, however, they lack fine-grained focus on local regions or objects. Inspired by (Shtedritski et al., 2023) which creates local focus with visual prompts; we use pre-extracted tube boxes as visual prompts to create candidate instances, rather than cropping out the region feature from the entire frame-level feature map.

Formally, we generalize video prompting as an operation to augment the video frames $\{\boldsymbol{v}_t\}$ with pre-extracted tube boxes $\mathcal{P}_i$ and the candidate instance is denoted as the prompted video frames $\tilde{\mathcal{P}}_i = \mathrm{Ops}(\mathcal{V}, \mathcal{P}_i) = \{\mathrm{Ops}(\boldsymbol{v}_t, \boldsymbol{b}_t^i)\}_{t=t_s}^{t_e}$. The generalized video prompting $\mathrm{Ops}(\cdot)$ can be drawing bounding boxes / drawing circles / drawing arrows / applying masks, cropping or even erasing. Notably, overlaying various visual markers on the image input is to create emphasis without loss of information while cropping or erasing is to discard certain information with a purpose. Since modifications are on input frames, we need to augment the textual input correspondingly to match the visual prompting. Specifically, textual prompt is inserted after the referred subject entity with the following template:

$$\mathrm{T}(\mathcal{S}) = \{\boldsymbol{s}_{\mathrm{subj}}\}\ \{\boldsymbol{s}_{\mathrm{prompt}}\}\ \{\boldsymbol{s}_{\mathrm{context}}\} \quad (3)$$

$\{\boldsymbol{s}_{\mathrm{subj}}\}$ is the subject entity and its attributive tokens and $\{\boldsymbol{s}_{\mathrm{context}}\}$ is the rest of the query sentence tokens, which can be obtained with off-the-shelf language parsing tools such as (Gardner et al., 2017). The text prompt $\{\boldsymbol{s}_{\mathrm{prompt}}\}$ inserted should be in accordance with the corresponding video prompt, for example, $\{\boldsymbol{s}_{\mathrm{prompt}}\} = \textit{in highlighted region}$ for brightness contrast adjustment over the candidate region, $\{\boldsymbol{s}_{\mathrm{prompt}}\} = \textit{in red circle}$ for a red circle, $\{\boldsymbol{s}_{\mathrm{prompt}}\} = \textit{pointed by red arrow}$ for an arrow marker. Note that operations such as cropping and erasing do not have corresponding textual prompts. In this paper, we experimented with several types of video prompt forms. For the cropping and erasing operation, we use the text query as is. Finally, we calculate the similarity score between candidate instance $\tilde{\mathcal{P}}_i$ and the prompted query text $\mathrm{T}(\mathcal{S})$ to determine which candidate is the best match:

$$i^* = \arg\max_i \boldsymbol{\phi}(\tilde{\mathcal{P}}_i, \mathrm{T}(\mathcal{S})) \quad (4)$$

Contrastive Video-Text Prompting (CVTP) framework Intuitively, applying textual prompt on subject word is equivalent to adding an additional attributive clause to the referent. This prompted text creates factual contradiction when the corresponding video prompt is landed on the incorrect candidate object tube, which encourages a lower similarity score. On the contrary, the prompted query fully aligns with the video frames that is prompted with the correct candidate boxes, thus encourages a high matching score. Furthermore, when erasing as a prompt is applied upon the correct entity, the similarity score should be low; but if we erase an entity that is irrelevant to the query, we would have even higher similarity scores simply because we removed some interference information.

Enlightened by above observation, we propose CVTP framework, as illustrated in Fig. 2. Concretely, for each pre-extracted tube boxes $\mathcal{P}_i$, we construct VP candidate instance $\tilde{\mathcal{P}}_i = \{\mathrm{Mark}(\boldsymbol{v}_t, \boldsymbol{b}_t^i)\}_{t=t_s}^{t_e}$ and its CVP counterpart $\bar{\mathcal{P}}_i = \{\mathrm{Erase}(\boldsymbol{v}_t, \boldsymbol{b}_t^i)\}_{t=t_s}^{t_e}$, where $\mathrm{Mark}(\cdot)$ and $\mathrm{Erase}(\cdot)$ indicates applying Video Prompting by marking and Contrastive Video Prompting by erasing, respectively. As such, Eq. 4 is updated as:

$$i^* = \arg\max_i \Big(\boldsymbol{\phi}(\tilde{\mathcal{P}}_i, \mathrm{T}(\mathcal{S})) - \boldsymbol{\phi}(\bar{\mathcal{P}}_i, \mathcal{S})\Big) \quad (5)$$

Video-Text Prompting Interactor As an implementation of the ranking function $\boldsymbol{\phi}(\cdot)$, we pro-

pose a novel VTP Interactor to model the cross-modal interaction, as well as temporal reasoning. Specifically, given a prompted candidate instance $\tilde{\mathcal{P}} \in \mathbb{R}^{T\times H\times W\times 3}$ and a query text $\mathcal{S}$; we extract their corresponding modal specific feature $\boldsymbol{V} \in \mathbb{R}^{T\times d}$ and $\boldsymbol{Q} \in \mathbb{R}^{(l+1)\times d}$ with visual and textual encoders of the pre-trained CLIP model, respectively. T, H and W is the shape of the input video tensor, l is the number of text tokens and we pre-pend one extra $[CLS]$ token to the query $\mathcal{S}$ for a sentence-level representation. d is the dimension of the latent feature. A light-weight transformer encoder is designed to model the temporal interaction between video frames with the self-attention layers. Given the fact that not all frames from the candidate tube are equally relevant to the query, we need to predict a temporal mask to filter the frame feature depending on the query. Before that, a transformer decoder is needed to model the cross-modal reasoning and perform the feature fusion:

$$\tilde{\boldsymbol{V}} = \boldsymbol{V} + \mathrm{FFN}\Big(\boldsymbol{V} + \mathrm{SelfAttn}(\boldsymbol{V}, \boldsymbol{V})\Big) \quad (6)$$

$$\tilde{\boldsymbol{Q}} = \boldsymbol{Q} + \mathrm{FFN}\Big(\boldsymbol{Q} + \mathrm{CrossAttn}(\boldsymbol{Q}, \tilde{\boldsymbol{V}})\Big) \quad (7)$$

where $\mathrm{FFN}(\cdot)$ is the Feed Forward Network, $\mathrm{SelfAttn}(\cdot)$ and $\mathrm{CrossAttn}(\cdot)$ is the multi-head self-attention and multi-head cross-attention layer, respectively.

We take the representation vector $\boldsymbol{h}_{cls}$ for the token $[CLS]$ from the decoder output $\boldsymbol{Q}$. A temporal mask generator is designed to perform the frame filtering as well as temporal localization. Inspired by (Zheng et al., 2022), we implement a simple MLP to predict the center c and the width w of the temporal span with the feature vector $\boldsymbol{h}_{cls}$ from $\tilde{\boldsymbol{Q}}$: $c, w = \mathrm{MLP}(\boldsymbol{h}_{cls})$. A temporal mask $\boldsymbol{m} \in [0,1]^T$ is constructed with the predicted c and w, which is used to weight and pool the temporally interacted video feature $\tilde{\boldsymbol{V}}$. Unlike (Zheng et al., 2022) where the mask is set to be gaussian-like, our mask has a steeper transition slope. Finally the similarity score between the prompted candidate $\tilde{\mathcal{P}}$ and the query $\mathcal{S}$ is calculated:

$$\phi(\tilde{\mathcal{P}}, \mathcal{S}) = \mathrm{Sim}\Big(\mathrm{AvgPool}(\boldsymbol{m} \circ \tilde{\boldsymbol{V}}), \tilde{\boldsymbol{Q}}\Big) \quad (8)$$

where $\mathrm{Sim}(\cdot)$ is a feature matching or similarity function that can be implemented as an MLP, dot product or other similarity metric. In this paper we implement $\mathrm{Sim}(\cdot)$ as cosine similarity.

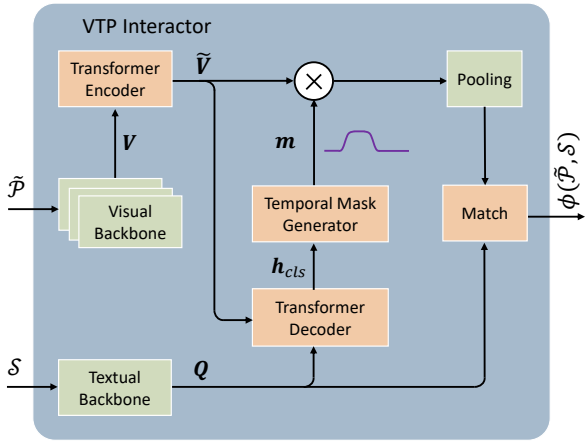


Figure 3: Video-Text Prompt(VTP) Interactor. The transformer encoder models interaction between video frames. The decoder models cross-modal interaction for the temporal mask generator to filter out the frames that are irrelevant to the query.

Training We train the VTP interactor with a ranking loss, which is commonly adopted in weakly-supervised setting. Specifically, with a semantically matched video clip $\mathcal{V}$ and query $\mathcal{S}$, unmatched video clip $\mathcal{V}'$ and query $\mathcal{S}'$ are randomly sampled from the same batch. The similarity score and ranking loss are calculated as follows:

$$s(\mathcal{V}, \mathcal{S}) = \max_i \Big(\phi(\tilde{\mathcal{P}}_i, \mathrm{T}(\mathcal{S})) - \phi(\bar{\mathcal{P}}_i, \mathcal{S})\Big) \quad (9)$$

$$\mathcal{L}_{rank} = \max[s(\mathcal{V}', \mathcal{S}) - s(\mathcal{V}, \mathcal{S}) + \delta, 0] \quad (10)$$

$$+ \max[s(\mathcal{V}, \mathcal{S}') - s(\mathcal{V}, \mathcal{S}) + \delta, 0] \quad (11)$$

where δ is a hyper-parameter for score margin. Additionally, to improve the effectiveness of the training process, we employ a candidates refinement step with the pre-trained CLIP encoder to rule out the candidates whose visual feature have a low similarity between with the subject of the query. This is to enforce the contrastive loss to focus only on difficult cases where the candidates share the same class.

4 Experiments

4.1 Datasets

VidSTG The dataset is proposed in (Zhang et al., 2020b) containing $44,808$ video samples paired with $99,943$ sentence queries. The length of the video clips range from 1 second to 2 minutes and each video sample contains 4.5 tube candidates on average. There are both humans and common objects for the referent of the queries

Methods	Declarative Sentences			Interrogative Sentences		
	m_vIoU	IoU@0.3	IoU@0.5	m_vIoU	IoU@0.3	IoU@0.5
Fully-Supervised (End-to-End)						
TubeDETR(Yang et al., 2022)	22.0	29.7	18.1	19.6	26.1	14.9
TubeDETR*(Yang et al., 2022)	30.4	42.5	28.2	25.7	35.7	23.2
CG-STVG(Gu et al., 2024)	34.0	47.7	33.1	29.0	40.5	27.5
Fully-Supervised (Two-Stage)						
GroundeR_T(Rohrbach et al., 2016)	9.78	11.04	4.09	9.32	11.39	3.24
STPR_T(Yamaguchi et al., 2017)	10.40	12.38	4.27	9.98	11.74	4.36
WSSTG_L(Chen et al., 2019b)	14.45	18.00	7.89	13.36	17.39	7.06
STGRN(Zhang et al., 2020b)	19.75	25.77	14.60	18.32	21.10	12.83
Weakly-Supervised						
AWGU(Chen et al., 2020)	8.96	7.86	3.10	8.57	6.84	2.88
Vis-Ctx(Shi et al., 2019)	9.34	7.32	3.34	8.69	7.18	2.91
WINNER(Li et al., 2023)	11.61	14.12	7.40	10.23	11.96	5.46
Ours						
VTP	16.12	19.39	13.28	11.13	12.2	8.0
CVTP	17.9	22.36	14.94	11.18	12.4	7.2

Table 1: Performance comparison on VidSTG(Zhang et al., 2020b). Notably, the end-to-end fully-supervised methods hold the state-of-the-art performance on this dataset thanks to the supervised training on object bounding-box regression. While the rest of the compared methods use cropped candidate feature obtained from pre-trained detectors, regardless of their training paradigm in terms of supervision. * indicates trained with extra-data.

and the class label are in accordance with the object detection dataset COCO (Lin et al., 2014). The dataset is constructed based on a Video Object Relation dataset(VidOR) (Shang et al., 2019) where both the visual content and query contain $< subject, predicate, object >$ triplet element. Additionally, the query could take the form of a question, referred to as interrogative sentences.

HC-STVG Proposed by (Tang et al., 2022), the dataset focuses on human activities and relationships where the referent is human in all samples. The dataset contains 5660 video-text sample pairs collected from movie scenes with a uniform duration of 20 seconds. On average, there are 5.3 candidate tubes in one video sample. Different from **VidSTG** where there is only one action in each query, samples in **HC-STVG** involves 2.3 actions on average.

4.2 Implementation

We extract and link frame-level object bounding boxes with off-the-shelf object detectors. However, different from previous works, we only use these boxes coordinates to insert our visual prompts; instead of directly extracting visual features from the detectors. We adopt the pre-trained CLIP model with ViT-L (Dosovitskiy et al., 2021) as our base encoders which remain frozen during our training. A two layer standard light-weight transformer encoder and decoder is designed with hidden dimension set to 256. An MLP is added to reduce the CLIP feature dimension to 256. For training, the margin δ is set to 0.2 and we use a batch size of 16 with a total training epoch of 10. The initial learning rate is set to $1e-4$.

4.3 Evaluation Metrics

We follow previous works (Su et al., 2021; Tang et al., 2022) by using m_vIoU and vIoU@R for evaluation. vIoU is a hybrid metric focusing on spatial grounding precision which is weighted by the overlapping time span between temporal prediction and ground truth, defined as $\text{vIoU} = \frac{1}{|\mathcal{S}_U|}\sum_{t\in\mathcal{S}_I}\text{IoU}(\boldsymbol{b}^t, \hat{\boldsymbol{b}}^t)$, where $\mathcal{S}_I$ and $\mathcal{S}_U$ is intersection and union between predicted and ground-truth frame span, respectively. vIoU@R reflects the percentage of test samples whose vIoU is larger than a threshold R, e.g. $R = 0.3$ and $R = 0.5$. Lastly m_vIoU is the mean of vIoU over all test set.

4.4 Performance Comparisons.

We compare our proposed prompting based methods VTP and CVTP with all existing weakly-supervised methods on VidSTG dataset in Table.

1. We also list some of the fully-supervised methods for a comprehensive study. The end-to-end supervised methods are able to refine their spatial grounding capability with the per-frame bounding-box annotation and this is to show what the SOTA performance is on this dataset. For a more meaningful comparison, we chose the two-staged ones because they all crop out candidate feature before they begin the reasoning or modeling process, regardless of their training paradigms. For supervised methods, GroundeR(Rohrbach et al., 2016), STPR(Yamaguchi et al., 2017) are only capable of grounding tubes spatially. Thus for the task of STVG, they employ a two-stage strategy by trimming the tube temporally with a trained temporal grounder such as L-Net(Chen et al., 2019a) or TALL(Gao et al., 2017). We append "T" and "L" to indicate the temporal grounding methods they adopted. Specially, WSSTG(Chen et al., 2019b) is trained under weak supervision for spatial grounding in trimmed video, however, its temporal localization part(L-Net) is an off-the-shelf model trained with temporal annotation.

We can see in Table 1 for the declarative sentences, our proposed prompt based VTP baseline has surpassed the SOTA in weakly-supervised methods by a tremendous margin on all metrics. Our Contrastive Video-Text Prompting(CVTP) method further expands this advantage. The IoU@0.5 percentage has doubled compared with the SOTA method WINNER(Li et al., 2023). Notably, our methods also surpasses those supervised methods with a two-stage strategy(GroundeR_T, STPR_T and WSSTG_L) by an impressive margin. While for the Interrogative sentences, where the subject word of the query is missing; which makes our prompt-based method less effective(this will be discussed in the Limitation section later). The proposed CVTP method still beats existing weakly-supervised methods on all metrics. Although the fully-supervised method STGRN(Zhang et al., 2020b) also conduct spatial and temporal grounding simultaneously, the performance gap between ours and STGRN is not as large as compared with others. Our proposed CVTP method even achieves a slightly higher retrieval percentage on IoU@0.5 metric for declarative sentences.

Evaluation results on HC-STVG(Tang et al., 2022) are shown in Table 2. The $*$ symbol with STGVT* and WSSTG* indicates the predicted tube is not temporally trimmed to produce this result. WSSTG_2D denotes 2D-TAN(Zhang et al., 2020a) is used for trimming. We note that although the STGVT(Tang et al., 2022) method chosen here leverages a transformer to model the temporal dependency and trained in a fully-supervised manner, its candidate tubes' features are still extracted by RoI pooling over local regions. Our methods are able to outperform it on one of the metric.

Methods	m_vIoU	IoU@0.3	IoU@0.5
Supervised			
WSSTG_T	13.37	19.95	7.33
WSSTG_2D	15.43	19.83	6.81
STGVT*	16.93	21.29	6.64
Weakly			
WSSTG*	12.96	16.23	4.35
AWGU	8.20	4.48	0.78
Vis-Ctx	9.76	6.81	1.03
WINNER	14.20	17.24	6.12
Ours			
VTP	16.15	18.48	6.65
CVTP	16.43	18.74	8.25

Table 2: Performance comparison on HC-STVG(Tang et al., 2022). Our proposed methods outperform all existing weakly-supervised methods. Some metrics even surpass supervised algorithms.

To summarize, our proposed prompting based methods are demonstrated to hold great advantage over those only utilize cropped regional feature for reasoning, this proves the importance of preserving context when extracting the candidate feature and this can be done via video-text prompting rather than isolating them from the global context.

4.5 Ablation Study

Ablation on Usage of Prompts First we would like to explore the effectiveness of different prompt used, as shown in Table 3. "V" refers to video prompting which is essential. "T" indicates text prompt added corresponding to the video prompting. "C" refers to the constructed Contrastive Video Prompted(CVP) counterpart used in our CVTP framework. As a retrieval setting, one of the pre-extract tube candidates is considered as the correct candidate during the testing stage who has the highest average IoU score with the ground truth tube. In this ablation, we additionally calculate the recall rate of our methods. The first row indicates randomly selecting a candidate from the pre-extracted tubes per testing video sample. Note that we do not randomly trim the tube hence this ablation focuses on spatial tube selection capability.

The performance gap between row 2 and row 4, the gap between row 3 and row 5 both highlight the importance of the text prompt: without the text prompt, the meaning of the video prompt is not explicitly specified hence resulting non-ideal performance. Similarly, the gap between row 2 and row 3 together with the gap between row 4 and row 5 highlight the effectiveness of the proposed contrastive ranking idea. Lastly, the proposed CVTP method in row 5 is threefold better than random in terms of recall accuracy.

Prompts			Metrics			
V	T	C	m_vIoU	IoU@0.3	IoU@0.5	Recall
			7.13	5.87	1.73	18.64
✓			13.12	14.11	5.28	41.77
✓		✓	14.02	16.32	6.39	44.73
✓	✓		16.15	18.48	6.65	53.45
✓	✓	✓	16.43	18.74	8.25	55.0

Table 3: Ablation on prompts used on HC-STVG dataset.

Ablation on Video Prompt Types Another ablation is conducted on the choice of video prompts, as shown in Table 4. As mentioned in Section 3.2, cropping can also be considered as a form of prompting operation. Here by "Highlight" prompting, we follow (Shtedritski et al., 2023) to adjust the brightness inside and outside of the candidate region where region inside the tube is brighter and the outside is darker. We also tested the video prompt as an arrow instead of a circle. The color used for both markers is red since it is both common and prominent in real world images. The results show that the arrow marker is slightly better than the circles (in this paper we stick to circles in all figures for better visualization), and cropping is the worst choice as it brings irreversible information loss.

Video Prompt	Metrics			
	m_vIoU	IoU@0.3	IoU@0.5	Recall
Crop	11.74	13.42	3.8	38.35
Highlight	14.8	17.01	6.48	47.58
Circle	15.89	18.05	7.48	52.76
Arrow	16.43	18.74	8.25	55.0

Table 4: Ablation on video prompt types on HC-STVG dataset.

A Feature Space View As mentioned in Section 3.2, by prompting the video and text input, we are creating emphasis over the prompted candidate. And this emphasis works positively towards correct candidate and negatively towards incorrect candidate. Here we provide a straight-forward illustration by showing the distance between visual and language feature prompted by different methods in Fig. 4. Specifically, for the candidates that we successfully retrieved, we extract their visual and language feature from our trained encoders and project them to 2-d space with t-SNE(van der Maaten and Hinton, 2008). The video and text feature with our prompts are in red and green, respectively. We also extract the candidates' feature by cropping in blue, and the original query feature without prompts in purple. As revealed in Fig. 4, with our proposed Video-Text Prompting, the average distance between video and language feature is much smaller than the distance between cropped visual feature and original query text.

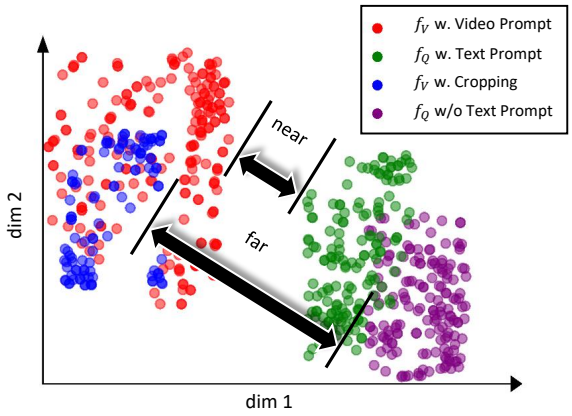


Figure 4: t-SNE visualization of the learnt visual and language feature for the retrieved candidates. Red and green are video and language feature with our proposed Video-Text Prompts; blue is cropped video feature and purple is original textual feature.

5 Conclusion

In conclusion, the proposed Video-Text Prompting (VTP) and Contrastive VTP (CVTP) effectively address the limitations of existing weakly-supervised STVG methods. By introducing video and text prompts instead of cropping object features, VTP preserves contextual information and enhances the representation of candidate features. Furthermore, CVTP leverages negative contrastive samples to improve the distinctiveness of correct candidates. Extensive experiments and ablations on multiple STVG datasets demonstrate the superiority of our approach, achieving significant performance im-

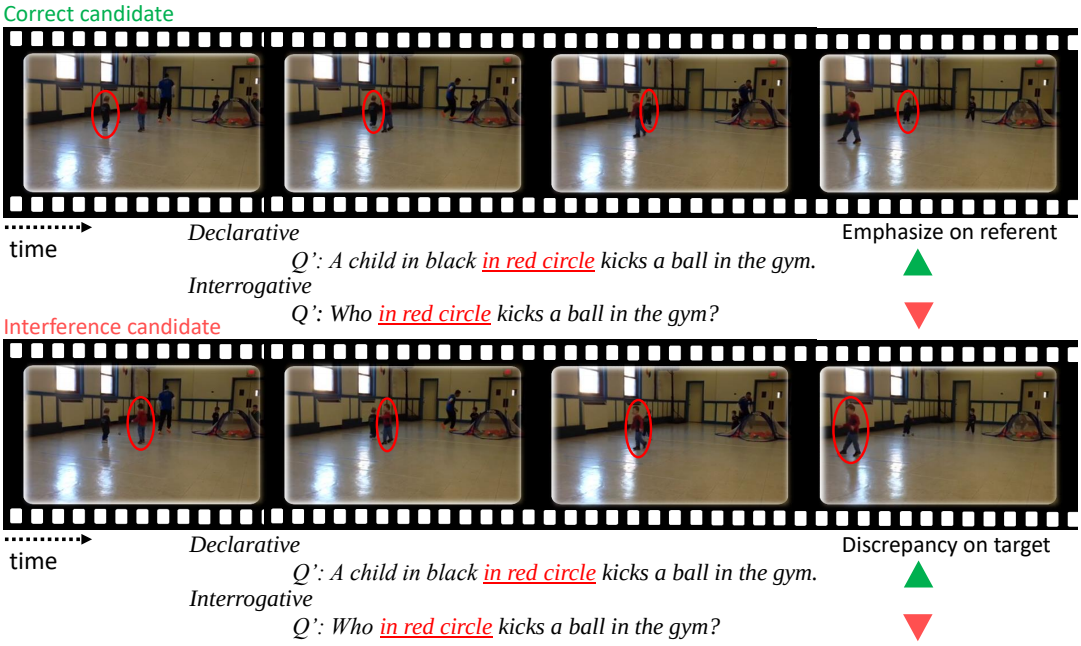


Figure 5: Illustration on the impact of different query forms to our methods.

provements over existing weakly-supervised methods. This highlights the potential of our methods in advancing the field of STVG without the need for densely annotated training data.

6 Limitations

6.1 Performance Upperbound

As mentioned in Section 3.1, in inference stage; we are picking the best matching tube candidate $\mathcal{P}^*$ given a finite set of candidate proposals $\{\mathcal{P}_i\}_{i=1}^{N_p}$ where N_p is the number of pre-extracted candidates. However, for all the metrics that are related to spatial precision, there is a gap between the best candidate $\mathcal{P}^*$ and the ground truth tube $\mathcal{P}_{gt}$. Thus in terms of numerical evaluation, the performance of our methods are upper-bounded by the metrics calculated between $\mathcal{P}^*$ adn $\mathcal{P}_{gt}$. Notably, this limitation applies to all weakly-supervised methods that formulate grounding as a retrieval problem.

6.2 The Form of Sentence Query

Proposed in (Zhang et al., 2020b), the VidSTG dataset includes declarative and interrogative sentences as the query texts. The former is of normal $< subject, predicate, object >$ triplet form; but in the latter one however, the $subject$ word is missing and the purpose is to force the model to reason like humans. From the results listed in Table 1 it can be seen that all methods suffer a performance drop with interrogative queries compared to those with declarative queries. We observe that the degradation with our model is notably higher than other methods. Intuitively, Mentioned in Section 3.2, by video prompting, our approach is essentially creating positive emphasis on correct candidate and negative discrepancies on incorrect ones and leverages the contrast between them. However with interrogative sentences, such emphasis is diluted. For example in Fig. 5, with a declarative sentence, both the emphasis and discrepancy enforced by the Video-Text Prompt method are strong; while for the interrogative sentence referring to the same entity, both the emphasis on correct candidate and the discrepancy on interference candidate drop. As a result, our proposed methods are more suitable for grounding with a normal declarative query.

6.3 VLM Encoder Reliance

Since our methods do not use the cropped feature from the object detectors, it's crucial for our visual and textual encoders to understand the prompts properly. As validated by (Shtedritski et al., 2023), VLMs trained on comprehensive web-scale vision-language data pairs such as the CLIP model is best suited for our model; for it can recognize and align the artificial Video-Text Prompts with reasonable confidence. However, we observe such capability is non-existent with visual encoders trained for specific tasks, such as object detection. Consequently, such encoders are not suitable to be incorporated in our framework. Nevertheless, the

trending paradigm for vision-language research is to leverage the broad world knowledge captured in foundation VLMs, we speculate such limitations will be less significant in the future.

7 Acknowledgements

This work is supported by Joey Tianyi Zhou's A*STAR SERC Central Research Fund (Use-Inspired Basic Research), A*STAR Centre for Frontier AI Research.

References

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Visual prompting: Modifying pixel space to adapt pre-trained models. *CoRR*, abs/2203.17274.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019a. Localizing natural language in videos. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8175–8182. AAAI Press.

Junwen Chen, Wentao Bao, and Yu Kong. 2020. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3789–3797. ACM.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. 2019b. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1884–1894. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. 2024. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. 2022. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, pages 105–124. Springer.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA,*

June 7-12, 2015, pages 3128–3137. IEEE Computer Society.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2024. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12861–12871.

Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. 2023. WINNER: weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23090–23099. IEEE.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. 2023. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23100–23109. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 817–834. Springer.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10444–10452. Computer Vision Foundation / IEEE.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does CLIP know about a red circle? visual prompt engineering for vlms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11953–11963. IEEE.

Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1513–1522. IEEE.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2022. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8238–8249.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Junyang Wu, Xianhang Li, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie. 2022. Unleashing the power of visual prompting at the pixel level. *CoRR*, abs/2212.10556.

Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Spatio-temporal person retrieval via natural language queries. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1462–1471. IEEE Computer Society.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16421–16432. IEEE.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. CPT: colorful prompt tuning for pre-trained vision-language models. *CoRR*, abs/2109.11797.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12870–12877. AAAI Press.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020b. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10665–10674. Computer Vision Foundation / IEEE.

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15534–15543. IEEE.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348.