VideoGrounding-DINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding

Syed Talal Wasim[1] Muzammal Naseer[1]
Salman Khan[1,2] Ming-Hsuan Yang[3,4] Fahad Shahbaz Khan[1,5]

[1]Mohamed bin Zayed University of AI [2]Australian National University
[3]University of California, Merced [4]Google Research [5]Linköping University

Abstract

Video grounding aims to localize a spatio-temporal section in a video corresponding to an input text query. This paper addresses a critical limitation in current video grounding methodologies by introducing an Open-Vocabulary Spatio-Temporal Video Grounding task. Unlike prevalent closed-set approaches that struggle with open-vocabulary scenarios due to limited training data and predefined vocabularies, our model leverages pre-trained representations from foundational spatial grounding models. This empowers it to effectively bridge the semantic gap between natural language and diverse visual content, achieving strong performance in closed-set and open-vocabulary settings. Our contributions include a novel spatio-temporal video grounding model, surpassing state-of-the-art results in closed-set evaluations on multiple datasets and demonstrating superior performance in open-vocabulary scenarios. Notably, the proposed model outperforms state-of-the-art methods in closed-set settings on VidSTG (Declarative and Interrogative) and HC-STVG (V1 and V2) datasets. Furthermore, in open-vocabulary evaluations on HC-STVG V1 and YouCook-Interactions, our model surpasses the recent best-performing models by 4.88 m_vIoU and 1.83% accuracy, demonstrating its efficacy in handling diverse linguistic and visual concepts for improved video understanding. Our codes will be publicly released.

1. Introduction

Spatio-temporal video grounding is pivotal in linking visual content with natural language descriptions, thus facilitating semantics interpretation within visual data. Prevailing approaches in video grounding such as TubeDETR [28], STCAT [9], and STVGFormer [11] focus mainly on supervised closed-set settings, where models are trained on specific datasets [24, 32] with predefined vocabulary and meticulously annotated data. While these current state-of-the-art methods excel in closed-set settings on datasets like VidSTG [32] and HC-STVG [24], their limited generalization beyond training dataset distributions poses a significant challenge. The relatively small scale and restricted sample variety in existing video datasets hinder models from adapting to unseen scenarios effectively.

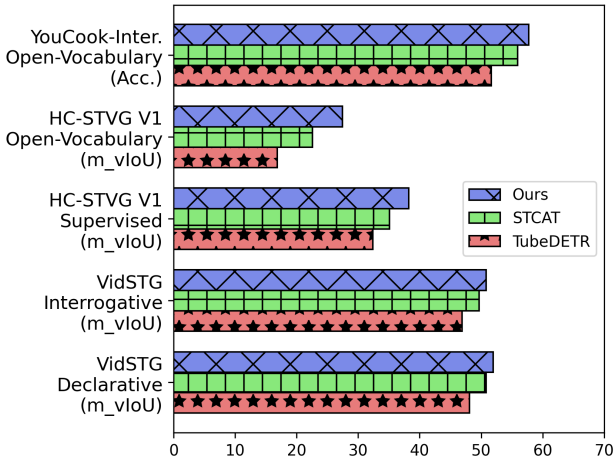

Figure 1. **Performance comparison** on conventional closed-set and open-vocabulary settings for the video grounding task. We compare our approach with TubeDETR [28] and STCAT [9] in supervised setting for VidSTG [32] declarative/interrogative and HC-STVG V1 [24], along with open-vocabulary evaluation on HC-STVG V1 and YouCook-Interactions [23] datasets.

Motivated by the inherent limitation of supervised closed-set approaches in terms of their restricted vocabulary, this paper investigates Open-Vocabulary Spatio-Temporal Video Grounding. Unlike conventional methodologies, this paradigm addresses challenges posed by the unrestricted diversity of language and visual concepts in the wild. The goal is to train on a set of base categories and generalize to unseen objects/actions based on the open-vocabulary nature of backbone models. This paper ex-

plores the challenges and opportunities inherent in open-vocabulary video grounding, laying the groundwork for more robust and versatile video understanding.

However, training an effective open-vocabulary video grounding model would require a large enough dataset with a rich set of natural language expressions and corresponding spatio-temporal localizations. Such an extensive dataset can allow the model to learn generalized visual and textual representations and handle out-of-distribution samples. However, video grounding datasets [24, 32] are quite limited in scale, e.g., VidSTG has only 5.4k training videos with 80.6k distinct sentences. In contrast, an image grounding model GLIP [10] is trained with $\sim$26.5M image-text pairs. Therefore, our research explores the following fundamental question: *How spatio-temporal video grounding models can achieve robust performance in both Closed-Set and Open-Vocabulary scenarios without requiring large-scale video annotations, ensuring effective generalization beyond training datasets?*

In addressing this question, we find inspiration in the accomplishments of foundational models specializing in spatial grounding [4, 8, 12, 14, 29]. These models undergo training on an extensive corpus of image-text data, enabling effective generalization to samples from a given target distribution. We aim to harness this pretrained representation to enhance our video grounding model. Our proposed solution is a spatio-temporal video grounding model adopting a DETR-like architecture enhanced by temporal aggregation modules. The spatial modules are initialized using the pretrained representation of a foundational image model [12]. Meanwhile, the image and text feature extractors remain frozen while the video-specific spatio-temporal adaptations are modeled via learnable adapter blocks. This approach is designed to preserve the nuanced representation of the foundational model, enhancing our model's ability to generalize effectively to novel samples.

A summary of our closed-set and open-vocabulary results is shown in Fig. 1, where the proposed approach excels in both settings by a clear margin. Our major contributions are summarized as follows:

- For the first time, we evaluate spatio-temporal video grounding models in an open-vocabulary setting on HC-STVG V1 [24] and YouCook-Interactions [23] benchmarks in a zero-shot manner. We outperform state-of-the-art methods TubeDETR [28] and STCAT [9] by 4.26 m_vIoU and 1.83% accuracy, respectively.
- By combining the strengths of spatial grounding models with complementary video-specific adapters, our approach consistently outperforms the previous state-of-the-art in closed-set setting on four benchmarks, i.e., VidSTG (Declarative) [32], VidSTG (Interrogative) [32], HC-STVG V1 [24] and HC-STVG V2 [24].

2. Related Work

Spatial Grounding Foundation Models: Recent literature introduces notable spatial grounding models. GLIP [10] unifies object detection and phrase grounding through a language-image pre-training model, leveraging extensive image-text pairs for semantic-rich representations. Grounding DINO [12] integrates language with a transformer-based detector to achieve an open-set detector, excelling in benchmarks like COCO and ODinW. Kosmos-1 [8] and Kosmos-2 [14] contribute Multimodal Large Language Models (MLLMs) with capabilities such as zero-shot and few-shot learning, language understanding, and multimodal tasks. Kosmos-2 [14] specifically integrates grounding into downstream applications, introducing GrIT, a large-scale dataset of Grounded Image-Text pairs. Shikra [4] addresses referential ability in MLLMs by handling spatial coordinates in inputs and outputs, showcasing promising performance in various vision-language tasks. Ferret [29] unifies referring and grounding in the LLM paradigm, achieving superior performance in classical referring and grounding tasks, excelling in region-based multimodal chatting and image description. Recently, GLaMM [16] allows pixel-level grounded conversations with an LLM, showcasing generalizability to several captioning and referring segmentation tasks. However, spatial methods cannot work for grounding objects in videos, a gap addressed by this work.

Spatio-Temporal Video Grounding: Several methods tackle the challenge of localizing objects in untrimmed videos based on query sentences. STVGBert [21] presents a one-stage visual-linguistic transformer for simultaneous spatial and temporal localization. TubeDETR [28] introduces a transformer-based architecture to model temporal, spatial, and multi-modal interactions efficiently. Augmented 2D-TAN [22] adopts a two-stage approach, enhancing the 2D-TAN with a temporal context-aware Bi-LSTM Aggregation Module. OMRN [33] addresses the challenge of unaligned data and multi-form sentences in spatio-temporal video grounding, proposing an object-aware multi-branch relation network for effective relation discovery. MMN [26] introduces a Mutual Matching Network as a metric-learning framework for temporal grounding, achieving competitive performance. STCAT [9] is an end-to-end one-stage framework addressing feature alignment and prediction inconsistency. Finally, STVGFormer [11] proposes an effective framework with static and dynamic branches for cross-modal understanding. While the above methods advance video grounding, their generalization to out-of-distribution and open-vocabulary samples is limited due to constrained video datasets [24, 32]. To address this issue, we utilize the generalized representation of spatial grounding foundation models [4, 8, 10, 12, 14, 29] trained on a large corpus of image-text data and can perform well on both closed-set and open-vocabulary evaluations.

3. Methodology

As discussed above, the current state-of-the-art spatio-temporal video grounding methods [9, 11, 21, 22, 24, 26, 28, 30, 32, 33] primarily evaluate in a supervised setting on the VidSTG [32] and HC-STVG [24] datasets. However, these methods lack the multimodal spatio-temporal understanding required to perform well on out-of-distribution samples [2]. Therefore, this work aims to achieve improved *open-vocabulary* performance while maintaining strong *closed-set* video-grounding performance.

We take inspiration from recent foundation models for spatial grounding [4, 8, 10, 12, 14, 29]. These models are trained on a large corpus of visual-textual data and hence, generalize well to unseen samples. We aim to leverage the strong generalization capabilities of such foundation models to achieve strong open-set spatio-temporal video grounding performance. Our proposed spatio-temporal video grounding method uses DETR-like [1] design, with temporal aggregation and adaptation modules for learning video-specific representations.

Below, we explain our proposed methodology. We formally define the spatio-temporal video grounding problem in Sec. 3.1. We then explain our architecture details in Sec. 3.2. We finally explain the loss formulation used to train the model and model initialization in Sec. 3.3.

3.1. Problem Definition

The spatio-temporal video grounding task involves localizing and recognizing objects and actions in a video sequence by integrating spatial and temporal information. In contrast to spatial grounding, which focuses on recognizing and localizing objects or actions within individual frames, spatio-temporal grounding extends this concept to include the temporal dimension. This means understanding where objects or actions are in each frame and how they evolve and move over time.

Consider a video $V \in \mathbb{R}^{T\times H\times W\times C}$ with T frames, $H \times W$ spatial resolution, and C channels, respectively, along with a text prompt P. The spatial grounding problem can be defined as the localization of one or more objects associated with the prompt P in a frame $V_t, t \in \{1, ..., T\}$ using a bounding box $B_i^t = (x_i^t, y_i^t, w_i^t, h_i^t)$, where x_i^t and y_i^t are the coordinates of the top-left corner, w_i^t and h_i^t are the width and height of the bounding box, $i \in \{1, ..., N\}$ is the object number for the frame t. The temporal grounding problem, on the other hand, involves understanding how objects or actions evolve over time. It aims at localizing the temporal interval (t_s, t_e) where the specific action/interaction happens in the entire temporal duration, where interval (t_s, t_e) indicates the start and end frame of the object occurrence within the total frames T ($1 \leq t_s < t_e \leq T$). Hence, the spatio-temporal grounding problem for object i associated with prompt P can be summarized as a set of spatio-temporal coordinates associated with the subset of frames where the object exists: $(x_i^t, y_i^t, w_i^t, h_i^t, t)$ and $t \in \{t_s, ..., t_e\}$. The interval $(t_s, t_e)|\{1 \leq t_s < t_e \leq T\}$ and is a subset of the total frames T.

3.2. Spatio-Temporal Video Grounding

Here, we explain our video grounding model in Fig. 2. As discussed earlier, we aim to design a spatio-temporal video grounding model that can perform well in closed-set and open-vocabulary settings. Strong open-vocabulary performance requires learning a rich visual/textual representation, which in turn requires a large amount of training data. Unfortunately, spatio-temporal video grounding datasets are quite limited in scale [24, 32], resulting in current video grounding methods failing in generalizing well to out-of-distribution samples because they lack the requisite strong visual/textual representation.

To solve this problem, our approach takes inspiration from recent spatial grounding methods [4, 8, 12, 14, 29], which have strong open-vocabulary performance thanks to the large image-text corpus they are trained on. We can utilize the generalized representations of these models to enrich the weaker representation of video-grounding approaches obtained from the limited number of training samples. Our approach aims to leverage the strong pre-trained representations of spatial grounding methods to achieve strong closed-set supervised and open-vocabulary video grounding performance. Our spatio-temporal video grounding approach is based on the state-of-the-art DETR-based [1] object detection framework DINO [31] and also borrows concepts of image-text alignment and grounding from Grounded Language-Image Pre-training (GLIP) [10] and Grounding DINO [12]. We extract initial features from backbone vision and text encoders θ_v and θ_p. Following that, we model inter-frame and intra-frame features and learn cross-modal visual/textual relations in the Cross-Modality Spatio-Temporal Encoder (Sec. 3.2.1). The result enriched cross-modal features are used to initialize queries for each frame (Sec. 3.2.2). These queries are then decoded to predict the bounding boxes per frame and the temporal grounding start/end frame by aggregating information across spatial/temporal dimensions and injecting information through cross-attention from enriched visual/textual context (Sec. 3.2.3).

Given the video V and text prompt P as described above, we obtain per-frame features F_v^0 and text features F_p^0 from vision and text encoders, θ_v and θ_p, respectively. The vision encoder is based on a Swin Transformer [13], and the text-encoder is defined as a BERT [6] model. Like other DETR-based detectors [31, 35], image features are extracted from different vision-encoder blocks at multiple scales. The features are then passed to the Cross-Modality Spatio-Temporal Encoder.

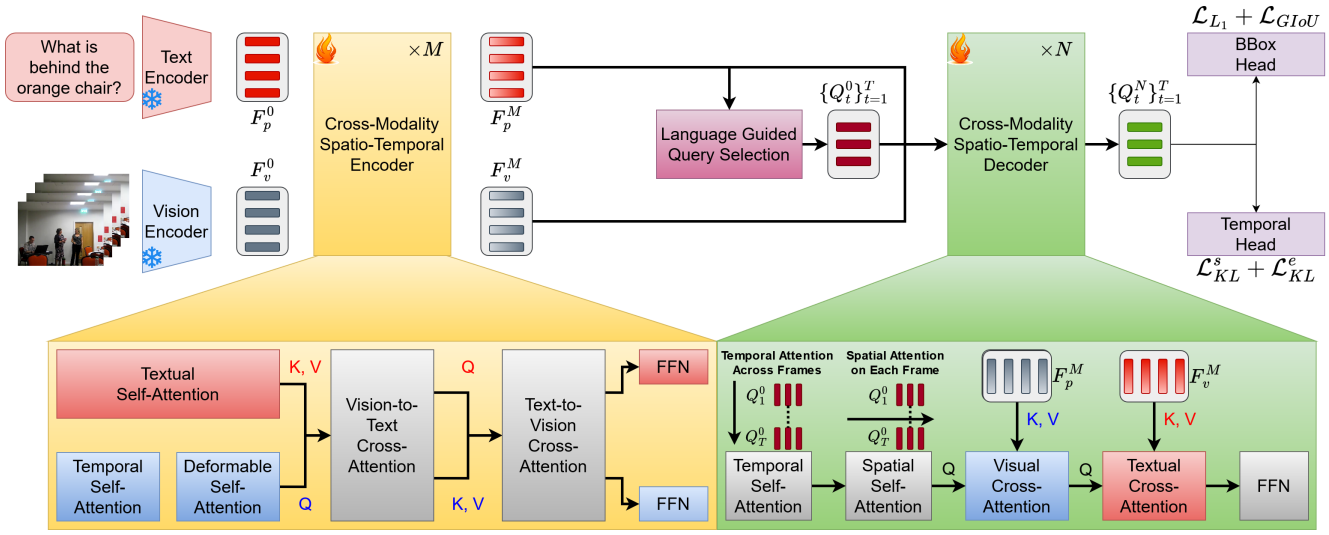

Figure 2. **Overall architecture:** We present our video grounding architecture. It consists of **vision and text encoders** that produce visual and textual features. A **cross-modality spatio-temporal encoder** which fuses information across spatial/temporal dimensions and visual/textual modalities. A **language guided query selction** module to initialize cross-modal queries. A **cross-modality spatio-temporal decoder** to decoder queries while fusing information from visual/textual features. And finally **two prediction heads** to predict the bounding boxes per frame and the temporal tube. Modules with (🔥) are trainable and those with (❄) are frozen.

3.2.1 Cross-Modality Spatio-Temporal Encoder

The initial vision and text features F_v^0 and F_p^0 neither contain any cross-modal information nor model the temporal relationship across frames. Therefore, we further encode the initial features through the Cross-Modality Spatio-Temporal Encoder to model the temporal information across frames and learn cross-modal features.

Each layer of the M layer encoder first applies a Multi-Head Self-Attention (MHSA) [25] to the visual features F_v along the temporal dimension, followed by a Deformable Attention (DA) [35] along the spatial dimension. This is done to model relations within frames and temporally across frames. Similarly, we apply a MHSA on the text features F_p. This is illustrated in Eq. 1.

$$\begin{aligned} F_v^{m\prime} &= \text{DA}_{spatial}^m(\text{MHSA}_{temporal}^m(F_v^{m-1})), \\ F_p^{m\prime} &= \text{MHSA}_p^m(F_p^{m-1}), \end{aligned} \tag{1}$$

where F_v^{m-1} and F_p^{m-1} are visual and textual input features to layer m, $F_v^{m\prime}$ and $F_p^{m\prime}$ are the intermediate visual and textual feature representation, and $\text{DA}_{spatial}^m$, $\text{MHSA}_{temporal}^m$ and MHSA_p^m are the spatial-deformable, temporal and textual attentions at layer $m \in \{1, ..., M\}$, respectively. Following initial spatial, temporal, and textual attentions, we fuse features across the visual and textual modalities, as done in GLIP [10].

More specifically, we calculate the joint visual-textual attention, $\text{Attn}_{\text{joint}}^{\text{m}}$, using projected intermediate features $F_v^{m\prime}$ and $F_p^{m\prime}$. This attention is then used alongside the intermediate features $F_v^{m\prime}$ and $F_p^{m\prime}$, to calculate the image-to-text and text-to-image cross-attentions as shown in Eq. 2 and Eq. 3.

$$\text{Attn}_{\text{joint}}^{\text{m}} = \left(\frac{proj_{q,v}^m(F_v^{m\prime}) proj_{q,p}^m(F_p^{m\prime})^T}{\sqrt{d^k}} \right), \tag{2}$$

where $proj_{q,v}^m$ and $proj_{q,p}^m$ are the query projections for the visual and textual features, respectively, at layer m.

$$\begin{aligned} F_v^m &= \text{FFN}_v^m(\text{softmax}(\text{Attn}_{\text{joint}}^{\text{m}}) proj_p^m(F_p^{m\prime}))), \\ F_p^m &= \text{FFN}_p^m(\text{softmax}(\text{Attn}_{\text{joint}}^{\text{m}\ T}) proj_v^m(F_v^{m\prime}))), \end{aligned} \tag{3}$$

where F_v^m and F_p^m are the final output features at layer m of the encoder, FFN_v^m and FFN_v^m are the visual and textual Feed Forward Networks (FFN) and $proj_v^m$ and $proj_p^m$ are the linear layers to project the visual and textual features. The final encoded features at layer $m = M$ are then utilized to initialize cross-modal queries.

3.2.2 Language-Guided Query Selection

This module is designed to select features more relevant to the input text as decoder queries for effective language-vision fusion. We combine a DETR/DINO [1, 31] style queries with a sinusoidal temporal positional encoding to the positional part of the queries. The sinusoidal positional encoding added to the positional part of the queries adds important contextual information regarding the sequence of frames, allowing for improved temporal correlation and

grounding [28]. The query selection module takes the encoder's visual and textual features as input and outputs num_query indices that correspond to the most relevant features for object detection per frame, $\{Q_t^0\}_{t=1}^T$, where Q_t^0 are the initial queries for the frame t. The module initializes the decoder queries using a combination of the selected indices and dynamic anchor boxes. The content part of the queries is set to be learnable during training, while the positional part is computed using the dynamic anchor boxes initialized using the encoder outputs. We also add a sinusoidal temporal positional encoding to the positional part of the queries.

3.2.3 Cross-Modality Spatio-Temporal Decoder

To decode the above queries into bounding box locations and temporal start/end tubes, we need to transform them into an output embedding, which can then be fed into prediction heads. The decoder allows for the queries to interact globally with others within a frame and across frames while utilizing the entire visual and textual features as context. Formally, the queries produced earlier are fed into a N layer decoder. Each layer starts with a temporal self-attention, a spatial self-attention followed by a visual cross-attention and textual cross-attention, and finally an FFN. This is represented in Eq. 4.

$$\begin{aligned} Q_t^{n\prime} &= \text{MHSA}^n_{spatial}(\text{MHSA}^n_{temporal}(Q_t^{n-1})), \\ Q_t^n &= \text{FFN}^n(\text{CA}^n_p(\text{CA}^n_v(Q_t^{n\prime}, F_v^M), F_p^M)), \end{aligned} \tag{4}$$

where Q_t^{n-1} are the input queries at layer $n \in \{1, ..., N\}$, $Q_t^{n\prime}$ are the intermediate queries after spatial and temporal attention at layer n, $Q_t^{n\prime}$ are the output queries at layer n, and CA_v^n and CA_p^n are the visual and textual cross-attentions at layer n. The cross-attentions are further elaborated in Eq. 5.

$$\begin{aligned} \text{CA}^n_v(Q_t^{n\prime}, F_v^M) &= \left(\frac{proj^n_{q,v}(Q_t^{n\prime}) proj^n_{k,v}(F_v^M)^T}{\sqrt{d^k}} proj^n_v(F_v^M)^T \right), \\ \text{CA}^n_p(\text{CA}^n_v, F_p^M) &= \left(\frac{proj^n_{q,p}(\text{CA}^n_v) proj^n_{k,p}(F_p^M)^T}{\sqrt{d^k}} proj^n_p(F_p^M)^T \right), \end{aligned} \tag{5}$$

where $proj^n_{q,v}$, $proj^n_{k,v}$ and $proj^n_v$ are the visual query, key and values projection for layer n and $proj^n_{q,p}$, $proj^n_{k,p}$ and $proj^n_p$ are the textual query, key and value projections. The final queries from the decoder at layer N, $\{Q_t^N\}_{t=1}^T$, are then used for prediction.

3.2.4 Prediction Heads

The decoder outputs refined queries per frame $\{Q_t^N\}_{t=1}^T$. We follow the standard DETR-like bounding box regression head implemented as a Multi-Layer Perceptron (MLP), which predicts bounding boxes $B_i^t = (x_i^t, y_i^t, w_i^t, h_i^t)$, per frame. To predict the temporal interval $(t_s, t_e)|\{1 \le t_s < t_e \le T\}$, we add a temporal grounding head, implemented as an MLP, alongside the bounding box regression head, similar to existing works like [9, 28]. The new head predicts the probabilities of the start $\tau_s \in [0,1]^T$ and ends $\tau_e \in [0,1]^T$ of the interval. During inference, the start and end interval $(t_s, t_e)|\{1 \le t_s < t_e \le T\}$ is computed by taking the maximum of the joint distribution of (τ_s, τ_e). Any invalid combinations with $t_e \le t_s$ are masked out.

3.3. Loss Function

To leverage the generalized pre-trained representation from spatial-grounding foundation models, we initialize all spatial modules and cross attentions from the Grounding DINO [12] spatial grounding model. To preserve this generalized representation while ensuring effective modeling of the downstream task, we freeze the Vision and Text Encoders θ_v and θ_p and fine-tune the remaining components.

During training, the model receives a batch of videos V with text prompt P. The ground-truth annotation contains the bounding box sequence $\{B_i^t\}_{t=t_s}^{t_e}$, and the corresponding start and end timestamps (t_s, t_e). For spatial grounding, we follow the standard loss formulation used in DETR-like [1, 31, 35], namely the L_1 loss, $\mathcal{L}_{L_1}$, and the Generalized Intersection over Union (GIoU) [17] loss, $\mathcal{L}_{GIoU}$. Formally, the spatial grounding loss, $\mathcal{L}_{spatial}$ is defined in Eq. 6.

$$\mathcal{L}_{spatial} = \lambda_{L_1}\mathcal{L}_{L_1}(\hat{B}, B) + \lambda_{GIoU}\mathcal{L}_{GIoU}(\hat{B}, B). \tag{6}$$

For temporal grounding, we follow [18, 21, 28] and generate two 1-dimensional gaussian heatmaps $\pi_s, \pi_e \in \mathcal{R}^T$, for the starting and ending positions. The temporal grounding loss is therefore defined in Eq. 7 as,

$$\mathcal{L}_{temporal} = \mathcal{L}^s_{KL}(\hat{\pi_s}, \pi_s) + \mathcal{L}^e_{KL}(\hat{\pi_e}, \pi_e), \tag{7}$$

where $\mathcal{L}^s_{KL}$ and $\mathcal{L}^e_{KL}$ are the KL divergence losses for the start and end distributions, respectively.

Note that the model outputs the bounding boxes and starting/ending distributions during inference. We determine the temporal grounding segment, (t_s, t_e), by taking the segment with the maximal joint start and end probability. Then, we consider the bounding boxes only within that tube for spatial grounding.

4. Results

4.1. Experimental Setup and Protocols

Below, we first briefly explain the implementation details (Sec. 4.1.1), followed by evaluation settings (Sec. 4.1.2), and datasets (Sec. 4.1.3) used in our work.

Table 1. Performance comparisons of the state-of-the-art on HC-STVG V1 [24] and YouCook-Interactions [23] in open-vocabulary setting.

Method	Pre-training	HC-STVG V1			YouCook-Interactions
		m_vIoU	vIoU@0.3	vIoU@0.5	Accuracy
TubeDETR *(CVPR'22)* [28]	VidSTG	16.84	22.32	9.22	51.63
STCAT *(NeurIPS'22)* [9]	VidSTG	22.58	32.14	20.83	55.90
VideoGrounding-DINO	VidSTG	**27.46**	**40.13**	**29.92**	**57.73**

4.1.1 Implementation Details

As discussed in the methodology (Sec. 3), we initialize the spatial modules in our model from the Grounding DINO [12] spatial grounding model and keep the vision and text encoders frozen. Our prediction heads for both spatial and temporal predictions are set to be 3-layer Multi-Layer Perceptrons (MLPs). We sample 128 frames during training and inference, resized to a resolution of 448 on the shorter side. We set both M and N to 6, and train the model with a batch size of 8 and learning rate of $1e^{-4}$, and weight decay if 10^{-4}. The number of epochs for VidSTG is set to 10, and for HC-STVG V1/V2 is set to 90.

4.1.2 Evaluation Settings

We evaluate our video grounding model in two settings, *Open-Vocabulary* and *Closed-Set Supervised*.

Open-Vocabulary Evaluation: In the open-vocabulary setting we train our model on the VidSTG [32] dataset and then evaluate on two different datasets, HC-STVG V1 [24] and YouCook-Interactions [23] to understand how well the model generalizes to new distributions. The reason for choosing these two datasets is that the former provides a relatively minor distribution shift given the similar perspective/objects in the videos compared to the training dataset VidSTG. In contrast, the latter provides a major distribution shift with changes in perspective and annotated objects/interactions.

Closed-Set Supervised Evaluation: In the supervised evaluation setting, we train on the training set and evaluate each dataset's respective validation/testing set. This evaluation is conducted for three majorly used datasets in spatio-temporal video grounding, namely VidSTG [32], HC-STVG V1 [24] and HC-STVG V2 [24].

4.1.3 Datasets

We evaluate our approach and compare against the state-of-the-art in two settings: *Open-Vocabulary* and *Closed-Set Supervised*, across a total of four grounding datasets, namely: VidSTG [32], HCSTVG V1 [24], HCSTVG V2 [24], and YouCook-Interactions [23].

VidSTG: The VidSTG [32] dataset is derived from the VidOR [20] dataset, incorporating object relation annotations. It includes 99,943 video-text pairs, encompassing 44,808 declarative sentence queries and 55,135 interrogative sentence queries. The training, validation, and test sets consist of 80,684, 8,956, and 10,303 sentences and 5,436, 602, and 732 videos, respectively. VidSTG's text queries are confined to describing pre-defined object/relation categories in VidOR [20].

HC-STVG V1/V2: The HC-STVG datasets are sourced from movie scenes, each video clip spanning approximately 20 seconds. These datasets pose challenges in spatio-temporal grounding due to video clips featuring multiple individuals engaged in similar actions. HC-STVG V1 comprises 4,500 training and 1,160 testing video-text pairs. HC-STVG V2 expands HC-STVG V1, enhancing annotation quality with 10,131, 2,000, and 4,413 samples for training, validation, and testing, respectively. As HC-STVG V2's test set annotations are unavailable publicly, results are reported on the validation set.

YouCook-Interactions: The YouCook-Interactions [23] dataset serves as an expansion of the YouCook2 [34] dataset focused on cooking instructions. This extension includes bounding boxes for 6,000 carefully chosen frames, typically encompassing the hand and the tool specified in the corresponding sentence-level annotations. Our assessment revolves around examining models' spatial grounding capabilities using this dataset.

4.2. Experimental Results and Analysis

In this section, we present our results across the evaluation mentioned above settings (Sec. 4.1.2) and datasets (Sec. 4.1.3). We start with the *closed-set* evaluation in Sec. 4.2.2, followed by the *open-vocabulary* evaluation in Sec. 4.2.1.

4.2.1 Open-Vocabulary Evaluation

For open-vocabulary evaluation, we train on VidSTG [32] and present results HC-STVG V1 [24] and YouCook-Interactions [23]. The results are reported jointly in Tab. 1.

Results on HC-STVG V1: We report open-vocabulary evaluation on m_vIoU, vIoU@0.3, and vIoU@0.5. We achieve state-of-the-art performance over both TubeDETR [28] and STCAT [9]. We attribute this strong performance to our design, which leverages the strong pre-

Table 2. Performance comparisons of the state-of-the-art on the VidSTG [32] test set in closed-set supervised setting.

Method	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Factorized:								
GroundeR *(ECCV'16)* [19]+TALL *(ICCV'17)* [7]		9.78	11.04	4.09		9.32	11.39	3.24
STPR *(ICCV'17)* [27]+TALL *(ICCV'17)* [7]	34.63	10.40	12.38	4.27	33.73	9.98	11.74	4.36
WSSTG *(arXiv'19)* [5]+TALL *(ICCV'17)* [7]		11.36	14.63	5.91		10.65	13.90	5.32
GroundeR *(ECCV'16)* [19]+L-Net *(AAAI'19)* [3]		11.89	15.32	5.45		11.05	14.28	5.11
STPR *(ICCV'17)* [27]+L-Net *(AAAI'19)* [3]	40.86	12.93	16.27	5.68	39.79	11.94	14.73	5.27
WSSTG *(arXiv'19)* [5]+L-Net *(AAAI'19)* [3]		14.45	18.00	7.89		13.36	17.39	7.06
Two-Stage:								
STGRN *(CVPR'20)* [32]	48.47	19.75	25.77	14.60	46.98	18.32	21.10	12.83
STGVT *(TCSVT'21)* [24]	-	21.62	29.80	18.94	-	-	-	-
OMRN *(IJCAI'21)* [33]	50.73	23.11	32.61	16.42	49.19	20.63	28.35	14.11
One-Stage:								
STVGBert *(ICCV'21)* [21]	-	23.97	30.91	18.39	-	22.51	25.97	15.95
TubeDETR *(CVPR'22)* [28]	48.10	30.40	42.50	28.20	46.90	25.70	35.70	23.20
STCAT *(NeurIPS'22)* [9]	50.82	33.14	46.20	32.58	49.67	28.22	39.24	26.63
STVGFormer *(CVPR'23)* [11]	-	33.70	47.20	32.80	-	28.50	39.90	26.20
VideoGrounding-DINO	**51.97**	**34.67**	**48.11**	**33.96**	**50.83**	**29.89**	**41.03**	**27.58**

Table 3. Performance comparisons of the state-of-the-art on the HC-STVG V1 [24] test set in closed-set supervised setting.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
STGVT *(TCSVT'21)* [24]	18.15	26.81	9.48
STVGBert *(ICCV'21)* [21]	20.42	29.37	11.31
TubeDETR *(CVPR'22)* [28]	32.40	49.80	23.50
STCAT *(NeurIPS'22)* [9]	35.09	57.67	30.09
STVGFormer *(CVPR'23)* [11]	36.90	62.20	34.80
VideoGrounding-DINO	**38.25**	**62.47**	**36.14**

Table 4. Performance comparisons of the state-of-the-art on the HC-STVG V2 [24] val set in closed-set supervised setting.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
Yu *et al (arXiv'21)* [30]	30.00	-	-
Aug. 2D-TAN *(arXiv'21)* [22]	30.40	50.40	18.80
TubeDETR *(CVPR'22)* [28]	36.40	58.80	30.60
STVGFormer *(CVPR'23)* [11]	38.70	65.50	33.80
VideoGrounding-DINO	**39.88**	**67.13**	**34.49**

trained generalized features of a spatial grounding foundation model.

Results on YouCook-Interactions: We evaluate our method further on the YouCook-Interactions [23] dataset, reporting pointing game accuracy for spatial grounding. Our approach gains nearly 2% in accuracy over STCAT [9] and more than 6% compared to TubeDETR [28]. This further shows our strong generalization capabilities in the open-vocabulary setting.

4.2.2 Closed-Set Supervised Evaluation

We present closed-set evaluations across three datasets, VidSTG [32], HC-STVG V1 [24] and HC-STVG V2 [24].

Results on VidSTG: We present results on the VidSTG test set in closed-set setting in Tab. 2, reporting m_tIoU, m_vIoU, vIoU@0.3 and vIoU@0.5. The results show that our method achieves state-of-the-art performance in comparison to both *Two-Stage* and *One-Stage* methods. In particular, we achieve more than $1t\_IoU$ gain in temporal grounding over the previous best methods OMRN [33] (*One-Stage*) and STVGFormer [11] (*Two-Stage*), both for Declarative and Interrogative sentences. Similarly, for m_vIoU, vIoU@0.3 and vIoU@0.5, we achieve a more than 1 unit gain the state-of-the-art methods STVGFormer [11] and STCAT [9]. Note that our method uses a frozen visual and textual encoder. In contrast, those mentioned above previous state-of-the-art methods all train the entire encoder.

Results on HC-STVG V1: We present results on the HC-STVG V1 dataset in Tab. 3, reporting m_vIoU, vIoU@0.3 and vIoU@0.5. We achieve a nearly 1.5 unit gain in m_vIoU and vIoU@0.5 and a 1 unit gain in vIoU@0.3 over the previous best method STVGFormer [11]. This shows the consistent performance of our method on this dataset.

Results on HC-STVG V2: We present results on the HC-STVG V2 dataset in Tab. 4, reporting m_vIoU, vIoU@0.3 and vIoU@0.5. Our performance gain on HC-STVG V1 is reflected here as well, with a consistent performance gain on HC-STVG V2, in comparison to the SoTA [11, 28].

Table 5. Ablation on various design choices for our approach on the VidSTG [32] test set in closed-set supervised setting.

Method	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Naive Solution (Frozen Grounding DINO [12])	39.78	18.07	22.31	13.75	39.79	9.66	10.42	3.84
+ Decoder Temporal Aggregation	42.81	20.74	26.53	15.41	43.81	12.38	16.71	8.62
+ Encoder Temporal Aggregation	46.29	23.19	32.38	18.95	47.17	16.19	23.28	13.04
+ Finetuned Spatial Modules in Decoder	48.06	28.97	41.60	26.06	49.58	24.27	32.85	20.11
+ Finetuned Spatial Modules in Encoder	51.97	34.67	48.11	33.96	50.83	29.89	41.03	27.58

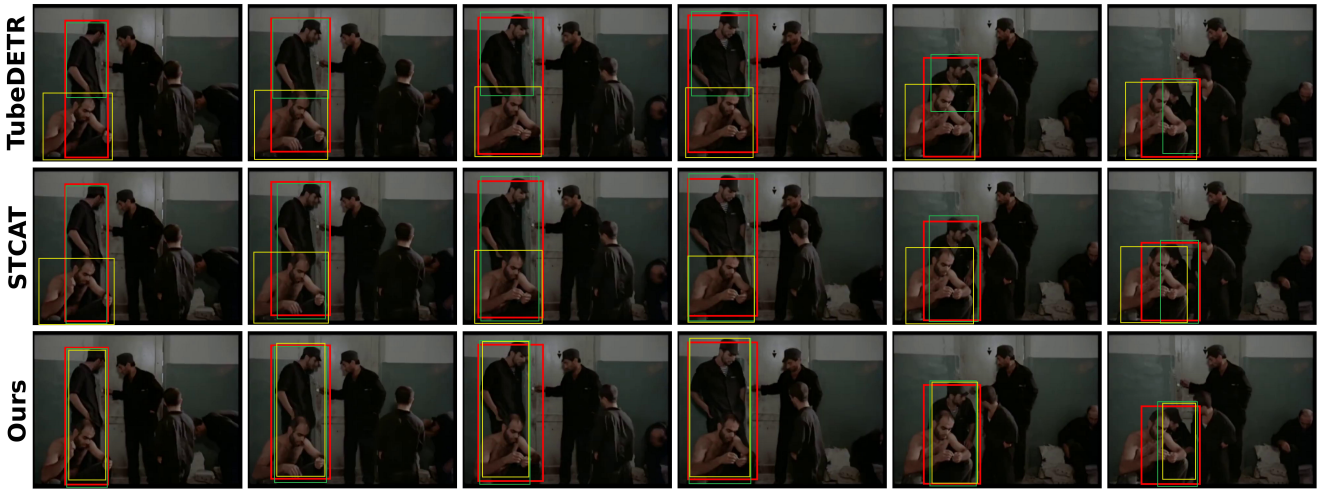

Figure 3. Sample visualization for video grounding result on HC-STVG V1 [24] for TubeDETR [28], STCAT [9] and ours with the prompt **The man behind the shirtless man turns and squats**. We show bounding boxes for Ground-truth, Closed-Set Supervised, and Open-Vocabulary results. Note how both TubeDETR and STCAT are close to the ground truth in the supervised setting (STCAT more so than TubeDETR), they cannot correctly ground the text properly in the open-vocabulary setting.

4.3. Ablative Analysis

We perform an ablative analysis of the various design choices for our model. In particular, we first evaluate a naive baseline where no additional temporal aggregators are added, and all pre-trained spatial modules are frozen in the encoder and decoder. This baseline is relatively weak in both temporal and spatial grounding. We next add temporal modules in first the decoder, and subsequently in the encoder. We find that it gives significant improvements in the temporal grounding, and additionally also improves the spatial grounding. Finally, we finetune the pre-trained spatial modules in both the decoder and the encoder, which provides a strong improvement in spatial grounding, alongside consistent improvement in temporal grounding.

4.4. Limitation

While our video grounding model excels in closed-set and open-vocabulary scenarios, it leverages image-text pre-trained models like Grounding DINO [12]. To enhance understanding in open-vocabulary settings, an extension to video-language pre-training on a larger and more diverse dataset, akin to CLIP [15], can help further boost performance. Building a video-language pre-training dataset with diverse natural language expressions and spatio-temporal localization is imperative, given the constraints of datasets like VidSTG [32] and HC-STVG [24].

5. Conclusion

This paper introduces an Open-Vocabulary Spatio-Temporal Video Grounding task, enhancing current closed-set methodologies by using pre-trained representations from spatial grounding models. The proposed model performs well in closed-set and open-vocabulary scenarios, surpassing state-of-the-art results in supervised setting on VidSTG and HC-STVG datasets, and outperforming recent models in open-vocabulary on HC-STVG V1 and YouCook-Interactions. Its architecture includes learnable adapter blocks for video-specific adaptation, bridging the semantic gap between natural language queries and visual content. This research addresses open-vocabulary challenges and explores achieving robust performance without extensive video annotations, paving the way for open-vocabulary video grounding.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3, 4, 5

[2] Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. What, when, and where? – self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions. *arXiv preprint arXiv:2303.16990*, 2023. 3

[3] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, , and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019. 7

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 3

[5] Zhenfang Chen, Lin Ma, Wenhan Luo, , and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019. 7

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. 3

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, , and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 7

[8] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2, 3

[9] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, 2022. 1, 2, 3, 5, 6, 7, 8

[10] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2, 3, 4

[11] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, 2023. 1, 2, 3, 7

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 5, 6, 8

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[14] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3

[15] 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. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 8

[16] Hanoona Rasheeda, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2

[17] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. In *CVPR*, 2019. 5

[18] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020. 5

[19] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, , and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 7

[20] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, 2019. 6

[21] Rui Su, Qian Yu, , and Dong Xu. Stvgbert: A visual- linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021. 2, 3, 5, 7

[22] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Xiang Li, and Wei-Shi Zheng. Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. *arXiv preprint arXiv:2106.10634*, 2021. 2, 3, 7

[23] Reuben Tan, Bryan A. Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos, 2021. 1, 2, 6, 7

[24] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1, 2, 3, 6, 7, 8

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[26] Zhenzhi Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, 2022. 2, 3

[27] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, , and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *ICCV*, 2017. 7

[28] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[29] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2, 3

[30] Yi Yu, Xinying Wang, Wei Hu, Xun Luo, and Cheng Li. 2rd place solutions in the hc-stvg track of person in context challenge 2021. *arXiv preprint arXiv:2106.07166*, 2021. 3, 7

[31] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3, 4, 5

[32] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[33] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *IJCAI*, 2021. 2, 3, 7

[34] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*, 2018. 6

[35] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3, 4, 5