This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CaptionFormer: Unified Segmentation, Tracking, and Captioning for Spatio-Temporal Objects

Gabriel Fiastre[1*] Antoine Yang[2] Cordelia Schmid[1]
[1]Inria, École Normale Supérieure, CNRS, PSL Research University [2]Google DeepMind
https://www.gabriel.fiastre.fr/captionformer/

Abstract

Dense Video Object Captioning (DVOC) is the task of jointly detecting, tracking, and captioning object trajectories in a video, requiring the ability to understand spatio-temporal details and describe them in natural language. Due to the complexity of the task and the high cost associated with manual annotation, previous approaches resort to training strategies with limited data, potentially leading to suboptimal performance. To circumvent this issue, we propose to generate captions about spatio-temporally localized entities leveraging a state-of-the-art VLM, and extend the LVIS and LV-VIS datasets with our synthetic captions (LVISCap and LV-VISCap). Moreover, we introduce an end-to-end model, CaptionFormer, capable of jointly detecting, segmenting, tracking and captioning object trajectories. With pretraining on LVISCap and LV-VISCap, CaptionFormer achieves state-of-the-art DVOC results on three existing benchmarks, VidSTG, VLN and BenSMOT.

1. Introduction

A fundamental aim of computer vision is to enable machines to understand videos with human-like acuity in perceiving and reasoning about the world. Recent advances have led to remarkable progress in both spatio-temporal localization [11, 26, 54, 55, 65, 77] and vision-language understanding [9, 14, 28, 30, 45, 58, 61, 68, 72, 74]. However, building vision-language models that can simultaneously reason about spatially localized objects and temporal dynamics of a complex scene remains a significant challenge, motivated by many real-world applications including autonomous driving [7, 33], human-computer interaction [4, 57], or video editing [27, 49]. Dense Video Object Captioning (DVOC) [81] serves as a key benchmark for this purpose, as it requires to jointly detect, track, and describe in natural language *all* visual entities in a video.

Manual annotation for such a fine-grained task is particularly expensive, leading to a scarcity of datasets with densely annotated object-level video descriptions. To tackle DVOC, prior work resorted to alternative training approaches: Zhou et al. [81] propose a disjoint training strategy, decomposing the problem into subtasks and training a model sequentially on datasets for each subtask. Choudhuri et al. [18] leverage the pretraining of multiple specialized models to alleviate the need for object-level annotations. Both methods allowed to perform DVOC while circumventing the need for costly annotations, but the lack of end-to-end training with densely annotated object-level supervision may lead to suboptimal performance.

Figure 1. Example of synthetic captions in our LV-VISCap dataset.

We propose to address this limitation by generating synthetic object-level annotations, motivated by the recent success of LLM-generated supervision [2, 43] and the growing visual capacities of Vision Language Models (VLMs) [3, 5, 8, 22, 38, 39, 59, 60]. To the best of our knowledge, our work is the first to generate localized, object-level captions for DVOC. To this end, we introduce a multi-modal prompting strategy leveraging a state-of-the-art VLM, and extend two segmentation datasets, LVIS [24] for images and LV-VIS [62] for videos, to be the first DVOC training sets with (mask, box, category, caption) annotations for all objects, dubbed LVISCap and LV-VISCap, see figure 1.

Using our generated datasets, we extend the DVOC task from detection to segmentation and train CaptionFormer, the first end-to-end model to jointly produce (mask, caption) pairs for all object trajectories in a video. We show that (i) our generated datasets, LVISCap and LV-VISCap, largely benefit CaptionFormer's DVOC performance, (ii) our CaptionFormer outperforms previous state-of-the-art on the VidSTG, VLN and BenSMOT benchmarks

*Corresponding author: gabriel.fiastre@inria.fr

and (iii) we can extend DVOC to segmentation.

Overall, our contributions can be summarized as follows:

1. We introduce a VLM-based method to generate synthetic object captions for videos, and extend the LVIS and LV-VIS datasets to be the first unified DVOC training set with object captions, boxes, and segmentation masks: LVISCap and LV-VISCap.
2. Using our unified generated data, we train CaptionFormer, the first end-to-end model to jointly detect, segment, track and caption objects in a video.
3. CaptionFormer achieves state-of-the-art DVOC results on the three existing benchmarks : VidSTG, VLN and BenSMOT.

The code and generated annotations are available at [1].

2. Related work

Open-Vocabulary Video Instance Segmentation (OV-VIS). The OV-VIS task aims to segment, track, and classify objects from an open set of categories in videos [23, 62], using datasets such as LV-VIS [62]. State-of-the-art methods [21, 23, 62] commonly use query-based approaches that classify objects by matching visual features with CLIP embeddings [52]. Methods like OVFormer [21] or BriVIS [17] improve this approach by better aligning visual queries with the CLIP space. Unlike these methods focused on CLIP feature matching for classification, our work explores describing objects in natural language [39].

Localized vision-language understanding. Going beyond pioneering vision-language tasks such as visual question answering [6] or image captioning [13], recent work has explored spatial understanding tasks that require localizing natural language queries in images. This includes referred expression segmentation [32, 70, 71], image grounding [51, 56], reasoning segmentation [35, 63], spatio-temporal video grounding [69, 78] and grounded visual question answering [36, 37, 67, 83]. While these tasks typically require localizing one or a few entities, dense captioning [29, 66] aims to spatially localize and describe in natural language *all* salient regions in images. Our work addresses the more challenging task of predicting *both* object trajectories and descriptions for *all* objects in a video.

Dense Video Object Captioning (DVOC). The DVOC task aims at jointly detecting, tracking, and describing the trajectory of all visual entities in a video. DVOC-DS [81] tackles this task by generating frame-wise object box proposals [10, 79] that are tracked [80] before feeding aggregated and cropped features to a generative image-to-text decoder [64]. To cope with the lack of DVOC annotations, the model is trained disjointly on various subtasks: object detection using COCO [42], image object-level captioning using Visual Genome [34], video scene-level captioning using SMiT [50] and video object tracking using AugCOCO [42]. OW-VisCapTor [18] extends the DVOC task to segmentation masks. Its architecture relies on an object abstractor using a prompt encoder and transformer blocks, an inter-query contrastive loss to ensure object queries are diverse, and an object-to-text abstractor that connects these queries to a frozen LLM, generating rich, object-centric captions for each detected instance. While proposing to extend DVOC to segmentation, OW-VisCapTor is hindered by the absence of paired (mask, caption) annotations, hence segmentation and DVOC are tackled in isolation using separate models. Closely related, SMOTer [41] extends multi-object tracking to object-level captioning, video-level captioning and object-relation predictions, and introduce a hand-annotated dataset, BenSMOT. However, BenSMOT focuses on humans only, whereas the standard DVOC task considers all visual entities in the video. In this work, we automatically generate DVOC datasets, LVISCap and LV-VISCap, and train CaptionFormer, a model that can end-to-end detect, segment, track and caption object trajectories in videos.

Vision-language data generation. A promising approach for intricate vision-language tasks is to generate visual annotations using Large Language Models (LLMs) or Vision Language Models (VLMs). LLaVA [43] leverages a LLM to generate conversations and detailed descriptions from paired image-text annotations [13]. This approach has been followed to generate large-scale instruction tuning data for video understanding [40, 47]. Recent research has focused on the generation of spatially grounded text-image data using LLMs/VLMs: Shikra [12] generates grounded QA pairs from Flickr30K Entities [51], LLaVA-Grounding [75] and GLaMM [53] generate grounded conversational data from datasets such as COCO [42]. Recently, GROVE [31] extended GLaMM to generate grounded, temporally consistent video captions, extending dense captioning to the video domain. These methods predominantly operate at the scene level and do not produce localized, object-level video descriptions. In contrast, we introduce a multi-modal prompting strategy to leverage VLMs into generating fine-grained captions for individual object trajectories across time. Recently, [73] proposed a related VLM-based approach to generate object-level video captions designed for long sentences referring segmentation. Differently, we focus on the task of Dense Video Object Captioning where no text prompt is given to the model.

3. Method

Dense Video Object Captioning (DVOC) [81] is the task of jointly detecting, tracking and captioning objects trajectories in a video, i.e. producing a (bounding box or mask, caption) pair for each object in a video at each timestamp they appear. Such densely annotated data is lacking for video, as there is currently no available training dataset including captions for all object trajectories in the videos. The spatio-temporal grounding dataset VidSTG [78] has been

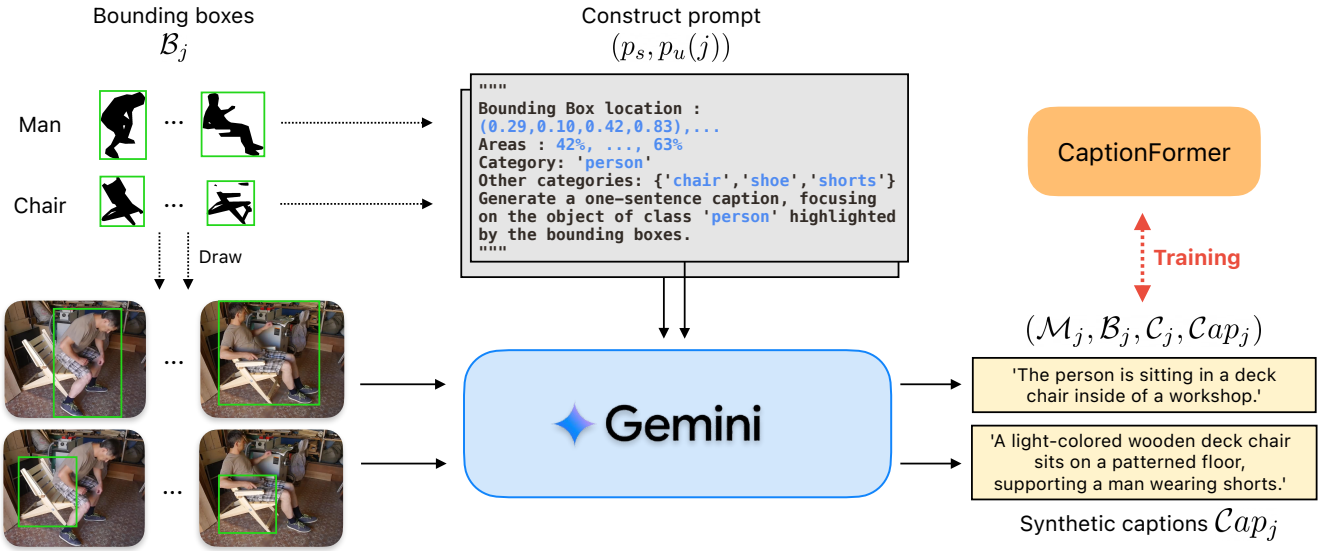


Figure 2. **Our CaptionFormer data annotation pipeline**: For each object, we extract its bounding boxes $\mathcal{B}_j$ from its annotated masks $\mathcal{M}_j$, and draw them in the video. The video with drawn bounding boxes $\hat{x}_j$, along with a prompt $(p_s, p_u(j))$ including additional information such as the label of the object to caption $\mathcal{C}_j$, the labels and bounding boxes of other objects in the image, is fed to a state-of-the-art VLM (Gemini 2.0 Flash [59]) to generate the object-level caption. p_s denotes the static prompt with general instructions and $p_u(j)$ the dynamic prompt with annotation information. We use this pipeline to generate the LVISCap and LV-VISCap datasets, used to train CaptionFormer.

repurposed to DVOC but only includes annotations for few objects per video and a limited number of frames per video.

We address the lack of data by introducing a strategy for synthetic DVOC data generation, allowing unified training of a DVOC model on (trajectory, caption) pairs for each object, as presented in Section 3.1. With this strategy, we extend the LV-VIS dataset [62], which includes both boxes and masks in their (trajectory, label) annotations for all objects, to a variant, dubbed LV-VISCap, which additionally includes synthetically generated captions for each trajectory. To enable end-to-end training on this rich data using segmentation masks, we build an architecture based on a state-of-the-art Open-Vocabulary Video Instance Segmentation (OV-VIS) model OVFormer [21], as described in Section 3.2.1. As OVFormer is designed for classification, we extend it with a captioning head [18], as explained in Section 3.2.2. Finally, we present in Section 3.3 the losses used to train our model.

Table 1. **Impact of the prompting strategy on caption quality**. Scores are given by an expert human evaluator from 0 to 2 (incorrect, partially correct, or correct) on a subset from the LV-VIS validation set, and scaled to 0-100 range. For the mask visual prompt experiments, we use our best prompt with either the object's bounding boxes or center point coordinates as a localization cue in the text prompt.

Visual prompt	Prompting method	Avg rating
bounding boxes	single frame	26.8
	+ multiple frames	27.1
	+ detailed instructions	29.5
	+ category labels	80.7
	+ bbox coordinates	83.1
	+ bbox area	84.3
	+ few shot examples	**85.1**
mask boundaries	center point coordinates	75.9
	bbox coordinates	77.1

3.1. DVOC data generation

We start from the LV-VIS dataset [62] which contains (segmentation masks, category) manual annotations for all objects and timestamps of the videos. To automatically collect DVOC data, one challenge is to generate accurate object-level captions for each trajectory. For this, we leverage a state-of-the-art VLM (Gemini 2.0 Flash [59]) and feed it with videos where the object to caption is marked with drawn bounding boxes, as illustrated in Figure 2.

Visual Prompt. Formally, let $x \in \mathbb{R}^{N \times H \times W \times 3}$ be a video clip of length N, with associated mask and category annotations for M objects in the video: $(\mathcal{M}_j, \mathcal{C}at_j)$, $j \in 1, ..M$. We first extract bounding boxes annotations from the ground-truth masks, $\mathcal{B}_j \in \mathbb{R}^{N \times 4}$, $j \in 1, ..., M$. We draw each object boxes on a separate copy of the video, and denote as $a \wedge b$ the operation of drawing box b on frame a. We obtain a visual prompt $\hat{x}_j^i$ for each object trajectory:

$$\hat{x}_j^i = x^i \wedge \mathcal{B}_j \quad \text{for } i \in 1, ..., N, \; j \in 1, ..., M \qquad (1)$$

Note that in practice, we subsample N to 4 uniformly sampled video frames as we found it produces representative enough visual content for the captioning task.

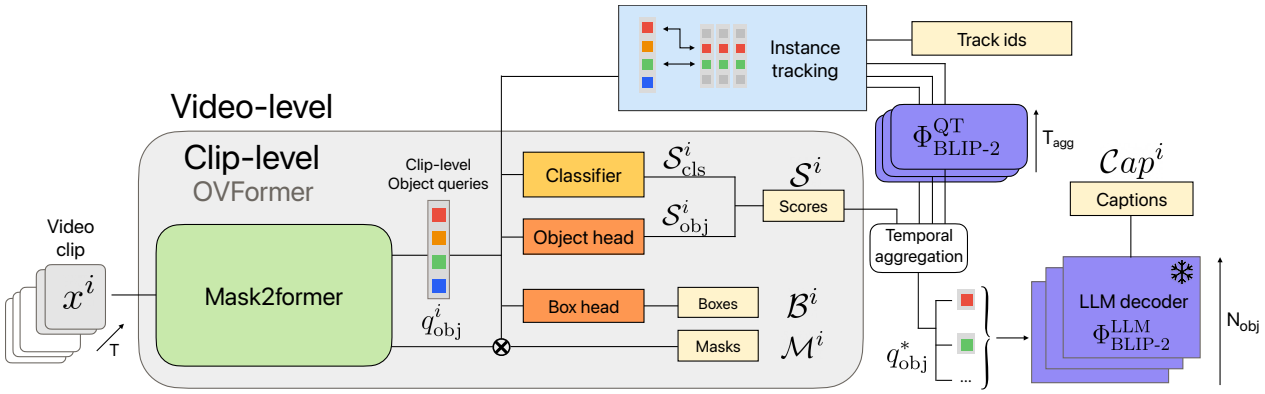


Figure 3. **Our CaptionFormer architecture** jointly segments and captions objects in videos. For each clip of T frames, we obtain N_{obj} clip-level object queries through Mask2Former [15], and yield associated score, mask and box predictions. At a video-level, we match the predicted object queries with the previously processed clips with the Hungarian bipartite matching algorithm, and perform instance tracking [82]. For each track, we sample T_{agg} clips uniformly across the video, and aggregate the tracked queries to obtain a single video query per object, which we can feed to the LLM captioning head [39], producing a single caption per trajectory.

Text prompts. In detail, the prompt we feed the VLM is composed of the previously described visual prompt $\hat{x}_j$, a system prompt p_s and a user prompt $p_u(j)$. The system prompt is static for all objects/videos and gives general instructions (e.g. "Generate a caption about a queried object highlighted with bounding boxes."), rules (e.g. "Do not mention the bounding boxes in the caption."), format, and an example. The user prompt $p_u(j)$ however, is constructed for each annotation and enriched with informations about the specific query to help the model. The user prompt encodes textual bounding box coordinates, area, category name of the queried object, and category names from other objects in the image. Passing information through different channels (visual prompt, text prompt) helps the model focusing on the queried object and being accurate when describing the scene. Also, this complementary information can lead the model to reason about the objects (e.g. area being small for an object of category 'elephant' implies the object is most likely part of the background). The different prompting cues have been ablated in Table 1. In particular, showing that including textual semantic and localization cues in the prompt helps the model to focus on the queried object and generate more accurate object-focused captions. We notice that using the segmentation masks as the visual prompt for the model results in less accurate object captions, which might result from a poorer alignment with the localization cues included in the text prompt. The full template details for the model prompt construction and a more detailed ablation are given in the supplementary material D.1 and B.

The LVISCap and LV-VISCap datasets. For each object j in the video, we prompt Gemini 2.0 Flash [59] with the visual and textual prompts ($\hat{x}_j$, p_s and $p_u(j)$) to output a synthetic caption $\mathcal{C}ap_j$. The full DVOC annotation is $(\mathcal{M}_j, \mathcal{B}_j, \mathcal{C}at_j, \mathcal{C}ap_j)$. We repeat this process for each object in each videos of LV-VIS, generating over 19.5k synthetic captions over 3.9k videos covering $1,020$ object classes, and with an average length of 13.9 words. We repeat this process with the LVIS dataset to support image pre-training, applying the same method as above described with each image considered a video of len $N = 1$, and obtain our DVOC training sets: LVISCap and LV-VISCap.

3.2. CaptionFormer architecture

To enable end-to-end training on the previously described DVOC data including segmentation masks, our architecture, illustrated in Figure 3, processes videos as N_{clip} clips of T consecutive frames each, and is composed with: (i) at the clip level, an instance segmentation and detection component, see Section 3.2.1 (ii) at the video-level, a tracking module and a captioning head, see Section 3.2.2.

3.2.1. Instance segmentation and detection

Background: OVFormer [21]. Our clip-level instance segmentation component is based on OVFormer [21], which we shortly describe next. On a high level, OVFormer augments Mask2Former [15] with a classification head to handle Open-Vocabulary Video Instance Segmentation. Mask2Former learns clip-level object queries with a transformer decoder which cross-attends to the features extracted with a visual backbone and are refined with a pixel decoder. Formally, each clip $x^i_{\text{clip}} \in \mathbb{R}^{T\times 3\times H\times W}$ is composed of T consecutive frames for images of resolution (H,W), and processed independently by the OVFormer model which outputs clip-level object queries q^i_{obj}, associated mask predictions $\mathcal{M}^i$, classification and objectness scores $\mathcal{S}^i_{\text{cls}}$ and $\mathcal{S}^i_{\text{obj}}$: $(q_{\text{obj}}, \mathcal{M}^i, \mathcal{S}^i_{\text{cls}}, \mathcal{S}^i_{\text{obj}}) = \Phi_{\text{OVFormer}}(x^i)$ where $q^i_{\text{obj}} \in \mathbb{R}^{N_{\text{obj}}\times D}$, $\mathcal{M}^i \in \mathbb{R}^{N_{\text{obj}}\times T\times H\times W}$, $S^i_{\text{cls}} \in \mathbb{R}^{N_{\text{clip}}\times N_{\text{obj}}\times N_{\text{cls}}}$ and $S^i_{\text{obj}} \in \mathbb{R}^{N_{\text{clip}}\times N_{\text{obj}}}$.

Detection head. We extend this segmentation module with detection by using a 4-layer MLP to generate boxes on top of the object queries q^i_{obj}: $\mathcal{B}^i = \text{BoxHead}(q^i_{\text{obj}}) \in \mathbb{R}^{N_{\text{obj}} \times T \times 4}$.

Confidence scores. Note that OVFormer [21] only computes class-aware query-wise confidence scores over the full video. However, for objects appearing only in a small subset of frames in the video this strategy could result in inaccurate scores. Moreover, for DVOC, we wish to avoid redundant predictions i.e. having two queries predicting a similar trajectory. Hence we additionally compute *class-agnostic* query-wise confidence scores *for each clip* $\mathcal{S}^i_{\text{cls}*}$, by taking the maximum classification score over all labels $c \in 1, ..., N_{\text{cls}}$ for each query and clip:

$$\mathcal{S}^i_{\text{cls}*} = \underset{c}{max}(S^i_{\text{cls}}(c)) \in \mathbb{R}^{N_{\text{clip}} \times N_{\text{obj}}} \tag{2}$$

Finally we derive the per-clip score $\mathcal{S}^i = \sqrt{\mathbf{S}^i_{\text{cls}*} \times \mathbf{S}^i_{\text{obj}}}$ which we use for filtering predictions below a threshold t_{thresh} at inference-time for every time step.

3.2.2. Instance tracking and captioning

Tracking module. To derive the output video-level trajectories from the clip-level predictions, we need to obtain a matching between the queries at time i and the queries at time $i+1$. For this, we perform tracking between the clips using the top-K enhanced query-matching module from Zhu et al. [82]. For each clip, this module keeps a memory bank containing the queries from the T_{match} previous clips. Among these, it identifies the K_{match} most matched clips, and computes the optimal assignment using the Hungarian bipartite matching algorithm. Using the K_{match} most matched clips helps reducing error propagation compared to the OVFormer [21] tracking module, which maps queries from time $i+1$ to time i directly. Notably, we can keep track of objects that disappear and re-appear in a video, whereas they are automatically lost using the OVFormer tracking module.

This method is referred to as semi-online tracking as we represent objects at a clip-level and associate between the clips in an online fashion. This offers the advantages of being flexible (fully-online for clips of length $T = 1$) and to arbitrate between using multi-frame information and memory constraints for long videos.

Captioning head. To caption tracked object trajectories, we adapt the captioning head from Choudhuri et al. [18] based on BLIP-2 [39]. The BLIP-2 decoder processes object queries one by one using masked-attention conditioned with the predicted masks, before projecting the resulting object query into the LLM space for caption prediction. However, for consistency and efficiency, we predict a single video-level caption per tracked object query, replacing clip-level prediction.

Query aggregation. Unlike previous works [18, 39], we extend the BLIP-2 decoder with temporal aggregation. Let's consider $\Phi^{\text{QT}}_{\text{BLIP2}}$ the BLIP-2 query transformer, $\Phi^{\text{LLM}}_{\text{BLIP2}}$ the LLM decoder, $q^i_{\text{obj}}(j) \in \mathbb{R}^{1 \times D}$ and $\mathcal{S}^i_{\text{obj}}(j) \in \mathbb{R}$ respectively the query and the detection score for object j from clip i. For each object, we aggregate the tracked queries over time after they are processed by the BLIP-2 query transformer, by sampling a set $\mathcal{I}_{\text{agg}}$ of T_{agg} clips uniformly across the video. We obtain a video query for each object j:

$$q^*_{\text{obj}}(j) = \sum_{i \in \mathcal{I}_{\text{agg}}} \mathcal{S}^i(j) \times \Phi^{\text{QT}}_{\text{BLIP-2}}(q^i_{\text{obj}}(j), \mathcal{M}^i_j) \tag{3}$$

We then compute the video captioning prediction of tracked object j from its aggregated query:

$$\mathcal{C}ap(j) = \Phi^{\text{LLM}}_{\text{BLIP2}}(q^*_{\text{obj}}(j)) \text{ for } j \in 1, ...N \tag{4}$$

The impact of using temporal aggregation is detailed in Section 4.2, and the choice of aggregation strategy is shown to be important in Section 4.2.3.

3.3. Model training

We train CaptionFormer with a combination of clip-level and video-level losses. For each clip, we predict masks/boxes, classification scores and captions, and derive the clip-level training objective as the following combination of supervised losses:

$$\mathcal{L}_{\text{clip-level}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{s}} + \mathcal{L}_{\text{cap}} \tag{5}$$

where $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{s}}$ are the VIS losses from Fang et al. [21], i.e. $\mathcal{L}_{\text{seg}} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}$, with $\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{ce}}$ the dice and cross-entropy segmentation losses respectively, and $\mathcal{L}_s = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}$ where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{obj}}$ are the cross-entropy losses for classification and objectness. We add detection and captioning losses $\mathcal{L}_{\text{det}} = \lambda_{l_1}\mathcal{L}_{l_1} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}$ where $\mathcal{L}_{l_1}$ and $\mathcal{L}_{\text{giou}}$ are detection losses from Yang et al. [69], and $\mathcal{L}_{\text{cap}} = \lambda_{\text{clip-lm}}\mathcal{L}_{\text{lm}}$, with $\mathcal{L}_{\text{lm}}$ the cross-entropy language modeling loss [81].

When including the temporal aggregation module for captioning, we train the captioning head at the video level, i.e. we predict a caption per object for the full video after the tracking is performed and each object-query has been augmented across time.

$$\mathcal{L}_{\text{video-level}} = \lambda_{\text{vid-lm}}\mathcal{L}_{\text{lm}} \tag{6}$$

CaptionFormer can be trained in a completely end-to-end manner. However, in practice, we train the model in two-stages for most of the experiments to alleviate memory constraints: we first train the segmentation/detection and classification model, then freeze it and tune the captioning head. The captioning head is trained either at the clip-level, or at the video-level when enabling the temporal aggregation module (i.e. $\lambda_{\text{clip-lm}} = 0$ or $\lambda_{\text{vid-lm}} = 0$). We set each loss respective weight when computed to $\lambda_{\text{dice}}, \lambda_{\text{ce}}, \lambda_{l_1} = 5$, $\lambda_{\text{giou}}, \lambda_{\text{cls}}, \lambda_{\text{obj}} = 2$, and ($\lambda_{\text{clip-lm}} = 1$ or $\lambda_{\text{vid-lm}} = 1$).

Table 2. **Impact of training with LVISCap and LV-VISCap and of the visual backbone on LV-VISCap DVOC.**

Backbone	LVIScap	LV-VIScap	CapA	DetA	AssA	CHOTA
SwinB	-	✓	37.9	48.1	89.5	54.7
	✓	-	30.0	34.3	**93.2**	45.8
	✓	✓	**43.6**	**54.3**	89.1	**59.5**
ResNet50	✓	✓	39.0	51.1	88.5	56.1

4. Experiments

4.1. Experimental setting

Datasets. LVIS [24] and **LV-VIS**[62] are large-vocabulary instance segmentation datasets, respectively for image and video. **LVISCap** and **LV-VISCap** denote our extensions of LVIS and LV-VIS (see Section 3.1), with respectively $1,2M/244k$ synthetic image object captions for the training/validation of LVISCap and $16k/3.7k$ synthetic video object captions for LV-VISCap (average of 5.4 objects per video). Note that in the absence of annotations on the test sets of LVIS and LV-VIS, we only extend the training and validation sets with captions, and use the validation set for evaluation.

Benchmarks. VidSTG[78] is a spatio-temporal video grounding dataset containing text descriptions serving as queries, which Zhou et al. [81] propose to use for DVOC evaluation. The repurposed training and validation sets count $5.4k/602$ videos for $15.1k/1.6k$ object trajectories with captions respectively. **Video Localized Narratives (VLN)** is similarly repurposed [81]. For each of the $5.1k$ training and $2.4k$ validation videos, the dataset contains 3 sparsely annotated frames with non exhaustive captions. **BenSMOT** contains bounding box trajectories and associated captions focusing exclusively on humans in videos, with an average of 2.2 instances per video. It counts $2.2k$ videos for training and $1k$ for evaluation. More details about the datasets are given in the supplementary material D.2.

Evaluation Metrics. Following prior work, we evaluate DVOC using the CHOTA metric introduced by Zhou et al. [81], which extends the widely used multi-object tracking HOTA metric [46]. CHOTA decomposes the DVOC task into three components: detection accuracy (DetA)[46], association accuracy (AssA)[46], and captioning accuracy (CapA)[81]. These components reflect the model's ability to (i) correctly localize objects, (ii) maintain their identity across frames, and (iii) generate accurate natural language descriptions. CHOTA is defined as the geometric mean of the three components: $\text{CHOTA} = \sqrt[3]{\text{DetA} \cdot \text{AssA} \cdot \text{CapA}}$. Matching between predicted and ground-truth trajectories is performed using Intersection-over-Union (IoU) thresholds, similarly to standard tracking evaluations [48]. To extend the metric to segmentation masks, we simply replace box-based IoU with mask-based IoU when computing CHOTA with segmentation masks.

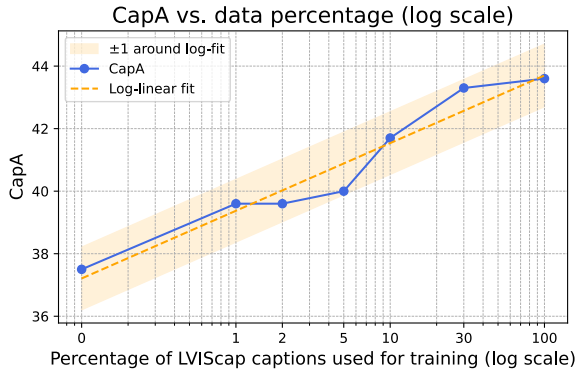


Figure 4. **Impact of the generated data scale on the CapA metric.** We train CaptionFormer on a varying percentage of LVISCap captions and finetune on LV-VISCap.

Implementation details. Following OVFormer [21] we use ResNet50 [25] and SwinBase [44] visual backbones; our Mask2former [16] transformer decoder has 11 layers, and our captioning head is based on the BLIP-2 [39] decoder with OPT-2.7B LLM [76]. For LV-VIS experiments we tune the model end-to-end with clip-level supervision only. For other experiments, we first train the segmentation/detection model, then freeze it and tune the captioning head. For VidSTG, VLN and BenSMOT experiments we use video-level tuning for captioning with temporal aggregation. For the largest dataset (COCO + LVIS) the optimization takes 2 days on 4 H100 GPUs. More implementation details and hyper-parameters for different experiments are given in the supplementary material D.3.

4.2. Results

4.2.1. Benefits of training on LVISCap and LV-VISCap

We first study the impact of integrating our generated datasets for DVOC training in Table 2. We train CaptionFormer on our synthetic video set, LV-VISCap, and progressively add LVISCap image pretraining. Results are reported on the LV-VISCap validation set. First, we observe that CaptionFormer achieves strong DVOC results with training only on LVISCap or LV-VISCap, demonstrating the effectiveness of our architecture. Importantly, combining both LVISCap and LV-VISCap leads to best results, showing the benefit of both our generated datasets. Moreover, we observe that our CaptionFormer is robust to the choice of the visual backbone. Note that the AssA score depends on the detections, hence a worse DetA score with lower recall but higher precision can make the tracking easier: this explains the higher AssA score from the model without LV-VIScap tuning. In Fig. 4, we show that the CapA performance is logarithmically correlated with the quantity of training captions, suggesting that generating more data with our approach might bring further improvements.

Table 3. **Comparison with the state of the art on VidSTG DVOC validation set.** All models except Gemini [19] are finetuned on VidSTG. For the OVFormer baseline, we use the official checkpoint to finetune on VidSTG and use predicted category label names as captions. "temp. agg." refers to including query temporal aggregation.

Method	Pretraining set	CapA	DetA	AssA	CHOTA
Gemini-2.5-flash[19]	-	6.4	0.9	47.2	6.6
OVFormer[21]	LVIS + LV-VIS	12.8	64.3	50.1	34.6
OW-VISCaptor [18]	COCO	43.9	60.1	54.0	53.0
CaptionFormer (ours)	COCO	44.3	65.1	70.2	58.7
DVOC-DS [81]	COCO + VG + SMIT + AugCOCO	39.7	65.8	70.4	56.9
CaptionFormer (ours)	COCO + VG + SMIT + AugCOCO	50.1	65.0	69.2	60.9
CaptionFormer (ours)	COCO + LVIScap + LV-VIScap	51.0	**66.8**	**71.0**	62.3
+ temp agg	COCO + LVIScap + LV-VIScap	**55.4**	**66.8**	**71.0**	**64.0**

Table 4. **Comparison with state of the art on the VLN DVOC validation set.** Mask loss refers to using the segmentation masks in the detection loss. All models are finetuned on VLN. "temp. agg." refers to including query temporal aggregation.

Method	Mask loss	CapA	DetA	AssA	CHOTA
DVOC-DS [81]	-	17.7	44.3	89.5	41.3
CaptionFormer (ours)	-	21.4	48.7	89.7	45.4
CaptionFormer (ours)	✓	22.9	**50.1**	**92.7**	47.4
+ temp agg	✓	**23.4**	**50.1**	**92.7**	**47.7**

Table 5. **Comparison on the BenSMOT validation set.** CapA, and thus CHOTA are not reported on this dataset. All models are finetuned on BenSMOT. "temp. agg." refers to including query temporal aggregation.

Method	DetA	AssA	CIDEr
SMOTer [41]	80.8	73.7	8.7
DVOC-DS [81]	90.8	**89.6**	25.4
CaptionFormer (ours)	**91.6**	87.5	39.9
+ temp agg	**91.6**	87.5	**42.6**

4.2.2. Comparison with the state of the art

We compare CaptionFormer to state-of-the-art DVOC methods following the standard evaluation protocol on three existing benchmarks : VidSTG in Table 3, VLN in Table 4 and BenSMOT in Table 5. DVOC-DS [81] reports results without pretraining, and with their disjoint training strategy, while OW-VISCaptor [18] leverages Mask2former pretrained on COCO for instance segmentation.

In Table 3, we include results with the same pretraining data as these methods and show the impact of using our data instead. Pretraining CaptionFormer on COCO yields better detection and tracking compared to OW-VISCaptor but comparable captioning performance due to the similar captioning head design. Including our data leads to an improvement on all metrics and especially captioning with a 6.7 CapA increase (15% relative improvement). When pretraining on the disjoint DVOC-DS set, we observe a substantial gain in the captioning metric due to the model design. Moreover, we show that using our pretraining sets results in a further performance improvement, while additionally allowing a unified, much faster training (2032 GPU hours [81] vs 208 for our approach). Moreover, our proposed approach can output segmentation masks unlike the other methods.

We include two additional baselines : (i) We prompt Gemini-2.5-flash [19] with the video frames and instruct it to generate the (trajectory, caption) pair for each object. Results are retrieved from textual answers with some tolerance on the output format. and (ii) a state of the art OV-VIS model, OVFormer [21], by using predicted category label names as predicted captions. Gemini [19] is unable to provide correctly formatted answers in 13.6% of the videos in the VidSTG validation set, and struggles to provide meaningful object regions, often generating near-stationary bounding boxes. This occasionally hinders any localized captioning capacity of the model, resulting in near 0 scores for both DetA and CapA. On the other hand, OVFormer [21] is able to provide meaningful object regions, but lacks captioning ability: using predicted category labels as captions can retrieve some semantic information but misses context, relevant actions and states of the described entities. Moreover, OVformer struggles to track objects due to the length of VidSTG videos. Overall, these results further show the gap between understanding and localization, highlighting the benefits of our unified DVOC approach.

We further evaluate CaptionFormer on VLN in Table 4 and BensMOT in Table 5. On both benchmarks, our approach improves the detection while the tracking remains competitive. Most important, the captioning metrics improve by a large margin (+5.2 CapA on VLN, +14.7 CIDEr on BenSMOT compared to state of the art). Additionally, our method is able to jointly segment objects, which further improves the performance as shown in Table 4.

Across all benchmarks, we show that including temporal aggregation further improves captioning performance by effectively merging information from multiple video clips, permitting the description of temporally extended actions.

Table 6. **Automatic vs manual annotations for evaluation on a subset from LV-VIScap validation** with 100 videos and 476 objects trajectories ; "**automatic**" and "**manual**" annotations stands for synthetic or human annotated captions. All models are trained on LVISCap and tuned on our LV-VISCap training set.

Annotation type	LVIScap captions	CapA	DetA	AssA	CHOTA
automatic	-	33.6	51.6	88.8	53.6
	✓	**41.4**	**52.1**	**90.2**	**58.0**
manual	-	22.9	51.6	88.8	47.1
	✓	**31.3**	**52.1**	**90.2**	**52.8**

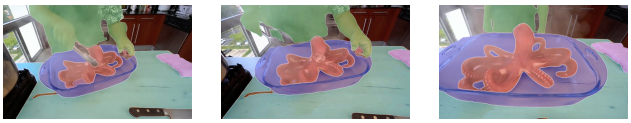
Table 8. **Impact of captioning on Video Instance Segmentation performance on LV-VIScap.** We train CaptionFormer respectively with, and without enabling the captioning loss on our LVIScap and LV-VIScap datasets.

Captioning loss	mAP
-	31.7
✓	**34.2**

We observe only a marginal improvement on VLN, likely due to the short video lengths of this dataset. Because we do not modify the detection and tracking models when adding temporal aggregation, the DetA and AssA scores are unchanged.

4.2.3. Additional ablations

Annotation bias. To evaluate the bias introduced by evaluating on LV-VISCap synthetic captions (see Table 2), we annotate a representative subset of our LV-VISCap datasets by hand and compare the impact of our LVISCap captions when evaluating CaptionFormer on *automatic* vs *manual* data. Our subset includes 100 videos from the LV-VISCap validation set with 476 object trajectories annotated by hand. The results are presented in Table 6. We observe that both the evaluation on automatic and manual data show a comparable improvement on the CapA and CHOTA metric when using LVISCap captions for training. This result confirms the importance of our synthetic captions for DVOC performance, and further shows that the bias introduced by evaluating on synthetic LV-VISCap captions is marginal.



A person wearing an apron is holding an octopus in a glass bowl.
The glass bowl is being filled with an octopus by a person's hand.
The octopus is being placed inside a glass bowl by a person.
A white towel is placed on the wooden table near a knife.
The wooden table is supporting a glass bowl filled with an octopus being handled by a person.

Figure 5. Qualitative example obtained with CaptionFormer on the LV-VIS dataset.

Table 7. **Impact of temporal aggregation on VidSTG validation** (with finetuning). All methods are pretrained on COCO + LVISCap + LV-VIScap. Multi-clip results are aggregated with weighted mean, except for † (arithmetic mean).

num clips	clip selection	CapA	CHOTA
1	best score	51.0	62.3
1	middle frame	46.9	60.6
4	uniform†	49.1	61.5
4	uniform	51.6	62.6
8	uniform	52.7	63.0
16	uniform	53.8	63.4
32	uniform	**55.4**	**64.0**

Temporal aggregation. We show the impact of temporal aggregation on DVOC performance in Table 7. Increasing the number of clips for aggregation consistently improves the captioning scores at the expense of higher training cost. Due to memory constraints we train with a maximum of 32 clips for aggregation on VidSTG. Weighting the clips for aggregation using the detection scores yields a clear improvement over the arithmetic mean. We also observe that performing captioning based on the single clip with best score performs relatively well, which highlights the limits of the complexity of the actions observable in the current benchmarks, e.g. VidSTG [78].

Captioning loss. We study the impact of the captioning loss on video instance segmentation (VIS) in Table 8, and observe that including our synthetic captions in the training improves mAP by a considerable margin. This improvement indicates that the captioning supervision yields a rich training signal that is propagated to the object queries. Moreover, this result hints at potential benefits of the DVOC task for large-vocabulary segmentation methods [62].

Qualitative example. In Figure 5, we show a qualitative sample produced by our CaptionFormer on the LV-VIS dataset. CaptionFormer effectively learns to jointly detect, track, segment and caption object trajectories in a video.

Additional results. We show additional qualitative examples in Supplementary A, and discuss failure cases and limitations of our method in Section C. Prompt ablation and the impact of the tracking module are presented in Section B and further implementation details in Section D.

5. Conclusion

We propose an approach to generate synthetic object-level captions using a state-of-the-art VLM and extend the LVIS and LV-VIS datasets with synthetic captions. We use the resulting LVISCap and LV-VISCap datasets to train CaptionFormer, a DVOC model that can simultaneously detect, segment, track, and caption objects throughout a video. With finetuning, CaptionFormer achieves state-of-the-art performance on the VidSTG, VLN and BenSMOT benchmarks, while extending the DVOC task to segmentation masks.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation AD011014323R2 made by GENCI. It was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "France 2030" program, reference ANR-23-IACL-0008 (PR[AI]RIE-PSAI project) and ANR project VideoPredict ANR-21-FAI1-0002-01. Cordelia Schmid would like to acknowledge the support by the Körber European Science Prize.

References

[1] CaptionFormer project webpage. https://www.gabriel.fiastre.fr/captionformer/. 2

[2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[4] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 2

[7] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 2024. 1

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 2

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

[12] 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

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation learning. In *ECCV*, 2020. 1

[15] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2Former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 4

[16] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6

[17] Zesen Cheng, Kehan Li, Li Hao, Peng Jin, Xiawu Zheng, Chang Liu, and Jie Chen. Aligning instance brownian bridge with texts for open-vocabulary video instance segmentation. In *AAAI*, 2025. 2

[18] Anwesa Choudhuri, Girish Chowdhary, and Alex Schwing. OW-VISCapTor: Abstractors for open-world video instance segmentation and captioning. In *NeurIPS*, 2024. 1, 2, 3, 5, 7, 6

[19] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[21] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *ECCV*, 2025. 2, 3, 4, 5, 6, 7

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[23] Pinxue Guo, Hao Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, and Wenqiang Zhang. OpenVIS: Open-vocabulary video instance segmentation. In *AAAI*, 2025. 2

[24] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 6, 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[27] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *ICLR*, 2024. 1

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[29] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 2

[30] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - Modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 1

[31] Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *ICCV*, 2025. 2

[32] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[33] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *CVPR*, 2019. 1

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2017. 2

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024. 2

[36] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018. 2

[37] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *ACL*, 2020. 2

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 2, 4, 5, 6

[40] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. In *CoRR*, 2023. 2

[41] Yunhao Li, Qin Li, Hao Wang, Xue Ma, Jiali Yao, Shaohua Dong, Heng Fan, and Libo Zhang. Beyond MOT: Semantic multi-object tracking. In *ECCV*, 2024. 2, 7

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2

[44] 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. 6

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1

[46] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. In *IJCV*, 2021. 6

[47] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 2

[48] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 6

[49] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1

[50] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021. 2

[51] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 6

[53] Hanoona Rasheed, 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. In *CVPR*, 2024. 2

[54] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *CVPR*, 2016. 1

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[56] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2

[57] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020. 1

[58] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019. 1

[59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk,

Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 3, 4

[60] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[61] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A neural image caption generator. In *CVPR*, 2015. 1

[62] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards Open-Vocabulary Video Instance Segmentation. In *ICCV*, 2023. 1, 2, 3, 6, 8, 5

[63] Junchi Wang and Lei Ke. LLM-Seg: Bridging image segmentation and large language model reasoning. In *CVPR*, 2024. 2

[64] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. In *TMLR*, 2022. 2

[65] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, 2017. 1

[66] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *ECCV*, 2024. 2

[67] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? Visually grounded video question answering. In *CVPR*, 2024. 2

[68] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just Ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 1

[69] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 2, 5

[70] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 2

[71] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2

[72] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1

[73] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv pre-print*, 2025. 2, 3

[74] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 1

[75] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. LLaVa-Grounding: Grounded visual chat with large multimodal models. In *ECCV*, 2024. 2

[76] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open Pre-trained Transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 6

[77] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *ECCV*, 2022. 1

[78] 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. 2, 6, 8, 1, 3

[79] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[80] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022. 2

[81] Xingyi Zhou, Anurag Arnab, Chen Sun, and Cordelia Schmid. Dense video object captioning from disjoint supervision. In *ICLR*, 2025. 1, 2, 5, 6, 7, 3

[82] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. CLIP-VIS: Adapting CLIP for Open-Vocabulary Video Instance Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 4, 5, 3

[83] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *CVPR*, 2016. 2