Aligning Instance Brownian Bridge with Texts for Open-Vocabulary Video Instance Segmentation

Zesen Cheng[1,3], **Kehan Li**[1,3], **Li Hao**[1,3], **Peng Jin**[1,3], **Xiawu Zheng**[5], **Chang Liu**[4] †, **Jie Chen**[1,2,3] †

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] Pengcheng Laboratory, Shenzhen, China
[3] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[4] Tsinghua University, Beijing, China
[5] Xiamen University, Xiamen, China

Abstract

Temporally locating objects with arbitrary class texts is the primary pursuit of open-vocabulary Video Instance Segmentation (VIS). Because of the insufficient vocabulary of video data, previous methods leverage image-text pretraining model for recognizing object instances by separately aligning each frame with class texts. As a result, the separation breaks the instance movement context of videos and requires a lot of inference overhead. To tackle these issues, we propose **B**ridge-**T**ext **A**lignment (BTA) to link frame-level instance representations as a Brownian Bridge. On one hand, we can calculate the global descriptor of a Brownian bridge for capturing instance dynamics, which enables extra considering temporal information rather than only static information of each frame for aligning with texts. On the other hand, according to the goal-conditioned property of Brownian bridge, we can estimate the middle frame features via the start and the end frame features so the global feature calculation of a Brownian bridge only needs to infer a few frames, which largely reduces inference overhead. We term our overall pipeline as BriVIS. Following training settings of previous works, BriVIS surpasses the SOTA (OV2Seg) by a clear margin. For example, on the challenging large-vocabulary datasets (BURST, LVVIS), BriVIS achieves 5.7 and 20.9 mAP, which exhibits +2.2∼+6.7 mAP improvement compared to OV2Seg. Furthermore, after training via BTA, using only the head and the tail frames for alignment improves the speed by 32% (2.77 → 1.88 s/iter) while just decreasing the performance by 0.2 mAP (21.1 → 20.9 mAP).

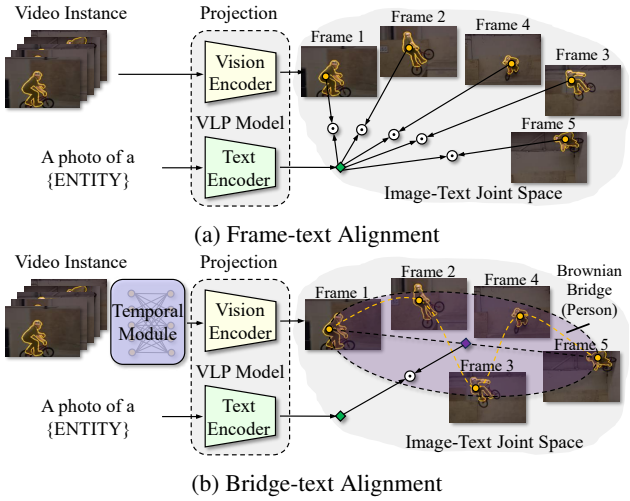

Figure 1: **The mechanical difference** between (a) frame-text and (b) bridge-text (ours) alignment. Previous methods recognize instances by integrating frame-text alignment results. Our method links frame-level instance features as a Brownian bridge and aligns the bridge descriptor to class texts, which can consider instance movement information when recognizing video instances. **Yellow circles**, **Green circles**, and **Diamond** denote frame-level instance features, class text features, and bridge descriptor.

Introduction

Video Instance Segmentation (VIS) aims to classify, segment, and track all object instances in the input videos (Yang, Fan, and Xu 2019). Based on the VIS task, the Open-Vocabulary Video Instance Segmentation (OVVIS) (Wang et al. 2023; Guo et al. 2023) aims to recognize any object categories for adapting real-world application scenarios. The development of image-text Vision-Language Pretraining (VLP) models (Radford et al. 2021; Jia et al. 2021; Yu et al. 2022; Yuan et al. 2021; Li et al. 2022a,b; Sun et al. 2023) largely boosts image-level open-vocabulary tasks (Gu et al. 2021; Huynh et al. 2022; Xu et al. 2022), attracting researchers to adapt image-text VLPs for open-vocabulary video tasks. Albeit directly constructing video-text VLP models can also advance open-vocabulary video tasks, it is not economical due to the expense of large-scale video-text pairs collection and annotation. Therefore, previous OVVIS models tend to leverage image-text VLP models for achieving open-vocabulary recognition.

To fit the image-text pretraining models' input modality, initial efforts (Wang et al. 2023; Guo et al. 2023) of OVVIS propose to depart video into frames and calculate instance-text alignment score frame by frame (Fig. 1a). Nevertheless, lacking the temporal modeling ability between frames, VLP models ignore how instance features evolve when aligning class text and instance features in a single frame. This results in suboptimal video instance recognition because the

† Corresponding Author

spatial-temporal context information contained in the video dynamics has been broadly demonstrated to be significant for instance semantic description (Tu et al. 2017; Ding et al. 2022; Hui et al. 2023). Moreover, frame-by-frame alignment between video and text has to infer a large number of frames, which requires a large amount of computation overhead. To solve these problems, we try to analyze instance movement and find that it contains a strong causal dependency during the process, i.e., the middle instance state can be easily inferred by observing the change between the start and end instance states. This agrees with the goal-conditioned property of the Brownian bridge and supports us to model the instance movement as a Brownian bridge. With the goal-conditioned property, the average of instance features on the start frame and the end frame, i.e., the bridge center, can serve as the descriptor of this Brownian bridge for roughly representing the whole instance movement. Therefore, when we align the bridge center to class texts, we can extra consider temporal information rather than only static information of each frame. Furthermore, because we only need to infer the start frame and the end frame, the computation overhead is largely reduced.

Specifically, we propose **B**ridge-**T**ext **A**lignment (**BTA**). Its conception is illustrated in Fig. 1b. The calculation flow of BTA contains three steps. Firstly, we constrain the Brownian bridge width by adopting hinge loss to modulate the distance between the head and tail instance features. Secondly, we derive a bridge-based contrastive objective to pull the middle frame features to follow the bridge distribution and push the middle frame features of other instances away. Finally, we adopt a contrastive objective to guide the bridge center close to the corresponding class text and away from irrelevant class texts. In total, we term our OVVIS scheme as **BriVIS**. Following previous work (Guo et al. 2023), we use Youtube-VIS 2019 (Yang, Fan, and Xu 2019) and COCO (Lin et al. 2014) as the training set. During the evaluation phase, we set large-vocabulary VIS datasets (BURST (Athar et al. 2023), LV-VIS (Wang et al. 2023)) as our benchmark. Our method achieves **5.7** mAP and **20.9** mAP on BURST and LVVIS, which is +2.2 mAP and +6.7 mAP better than previous state-of-the-art OV2Seg (Wang et al. 2023). This justifies that our Brownian bridge modeling provides more precise alignment with texts. Furthermore, we find that using only the first and the end frames of an instance for alignment improves the BriVIS inference speed by 32% (2.77 $\rightarrow$ 1.88 s/iter) while just decreasing the BriVIS performance by 0.2 mAP (21.1 $\rightarrow$ 20.9 mAP), which demonstrates that the goal-conditioned property can help OVVIS model reduce inference cost.

Related Work

Video Instance Segmentation

VIS is a comprehensive video recognition task, aims to segment, which aims to track, and classify all objects in videos (Yang, Fan, and Xu 2019). Numerous early endeavors provide their designs for accomplishing this task, e.g., one-stage methods (Han et al. 2022) and two-stage methods (Yang, Fan, and Xu 2019), but Transformer-based VIS methods gradually become the mainstream route because the attention architecture significantly improves the performance and the query-oriented design provides an elegant instance representation. VisTR (Wang et al. 2021) designs an end-to-end baseline to track the identical object across video via a simple instance query, which is the first attempt to extend DETR (Carion et al. 2020) to the VIS task. Follow-up works develop in two directions: online and offline. The online route departs video into frames or clips to process and focuses on how to link objects spanning the adjacent video frames or clips (Huang, Yu, and Anandkumar 2022; Heo et al. 2023; Zhang et al. 2023b). For example, MinVIS associates objects by using Hungarian algorithm (Kuhn 1955) to match queries between two adjacent frames. The offline route sets the whole video as a spatial-temporal volume to process and focuses on designing efficient and effective temporal interaction (Cheng et al. 2021; Heo et al. 2022; Li et al. 2023), e.g., IFC (Hwang et al. 2021) designs memory tokens while Seqformer (Wu et al. 2022a) proposes frame query decomposition. Although the offline route achieves more notable performance, the online route is more efficient, especially for long videos.

Open-Vocabulary Segmentation

Open-vocabulary image segmentation explores how to recognize any categories at pixel level. Earlier works (Zhao et al. 2017; Xian et al. 2019; Bucher et al. 2019) attempt to build a pixel-text alignment space by learning to align pixel embedding to word embeddings (Mikolov et al. 2013; Miller 1995) of class texts. With the rise of VLP models (Jia et al. 2021; Yu et al. 2022; Yuan et al. 2021; Li et al. 2022a,b; Sun et al. 2023) represented by CLIP (Radford et al. 2021), recent research focus shifts to explore how to adapt their superior image-level open-vocabulary ability to pixel-level. Part of the works (Ghiasi et al. 2022; Liang et al. 2023; Zhou, Loy, and Dai 2022) proposes to directly finetune or distill VLP for remedying this granularity gap, which requires vast segmentation data. Another part of the works (Zheng Ding 2023; Han et al. 2023) explore to reuse the original image-text alignment space of VLP. For example, SimpleBaseline (Xu et al. 2022) proposes to recognize image crops by a frozen CLIP, and SideAdapter (Xu et al. 2023) predicts attention biases to modulate self-attention of CLIP for mask-guided image-text alignment. In this paper, we mainly refer to the design philosophy of the latter part of open-vocabulary segmentation works and explore how to adapt VLP for spatial-temporal segmentation tasks.

Brownian Bridge Modeling

Brownian bridge (Revuz and Yor 2013) is a goal-goal-conditioned Gaussian stochastic process $B(t)$ whose probability distribution of each time step follows a Gaussian distribution and is conditioned by start state z_0 at $t = 0$ and end state z_T at $t = T$:

$$B(t) = \mathcal{N}((1 - \frac{t}{T})z_0 + \frac{t}{T}z_T, \frac{t(T-t)}{T}), \qquad (1)$$

where $z_t, t \in [0, T]$ is the middle state of the bridge. In Eq. 1, we can find that z_t is approximately the noisy linear inter-

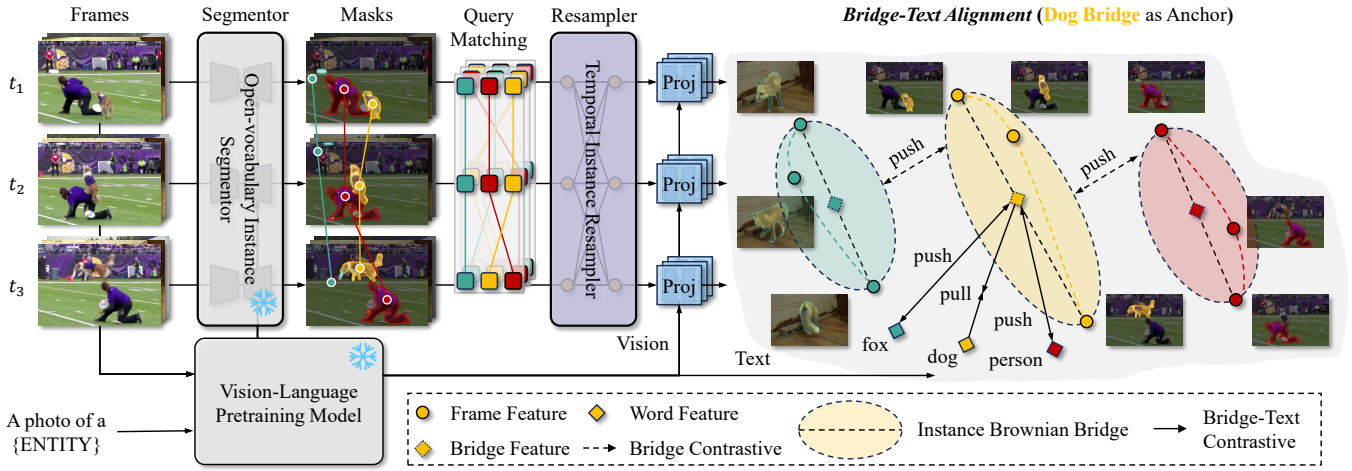

Figure 2: **The overall pipeline** of our BriVIS. Our main designs are BTA. It is used to link instances spanning different frames as a Brownian bridge and align them and text at bridge granularity.

polation of z_0 and z_T modulated by the time variable. The uncertainty gradually decreases to the lowest at the start or end side and increases to the highest at the center point.

Brownian bridge is a promising tool for modeling process-oriented problems. TC (Wang et al. 2022) proposed to generate text in a latent space with Brownian bridge dynamics. In this way, the middle text can follow local coherence and be modulated by start and end context, which is the pilot work to leverage the goal-conditioned nature of the Brownian bridge. Subsequently, the Brownian bridge is further explored in fine-grained video self-supervised learning (Zhang et al. 2023a) and dialogue generation (Wang, Lin, and Li 2023). In our work, the Brownian bridge is adopted to model temporal information based on frame-level instance features, which improves the alignment between the instance and the class text in the image-text semantic space when recognizing instances in videos.

Methods

Overall Pipeline

To provide a concrete and vivid description of our architecture, we illustrate our overall pipeline in Fig. 2.

Open-vocabulary Instance Segmentor. We choose Mask2Former (Cheng et al. 2022) as our basic segmentor to generate masks and queries for each frame. Mask2former is comprised of three parts: image backbone, pixel decoder, and query decoder. ❶ The video is first departed into frames for extracting backbone features $\{\boldsymbol{\mathcal{F}}_t^{b_1} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d_1}, \boldsymbol{\mathcal{F}}_t^{b_2} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d_2}, \boldsymbol{\mathcal{F}}_t^{b_3} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times d_3}\}$, where H, W, and d denote height, width, and feature dimension, respectively. ❷ Then a pixel decoder gradually integrates backbone features to generate high-quality multi-scale pixel features. Several layer features of CLIP model, meanwhile, are injected in three stages of pixel decoder to enhance the open-vocabulary ability:

$$\boldsymbol{\mathcal{F}}_t^{o_i} = \texttt{MSDefAttn}(\texttt{MLP}(\boldsymbol{\mathcal{F}}_t^{b_i}) + \texttt{MLP}(\boldsymbol{\mathcal{F}}_t^{c_i})), \quad (2)$$

where $\texttt{MSDefAttn}$ denotes multi-scale deformable transformer, $\boldsymbol{\mathcal{F}}_t^{o_i}$, $\boldsymbol{\mathcal{F}}_t^{b_i}$, and $\boldsymbol{\mathcal{F}}_t^{c_i}$ respectively denote the i-th stage open-vocabulary, backbone, and CLIP features of t-th frame. The shape of $\boldsymbol{\mathcal{F}}_t^{c_i}$ is processed to align $\boldsymbol{\mathcal{F}}_t^{b_i}$. The spatial shape of $\boldsymbol{\mathcal{F}}_t^{o_i}$ is equal to $\boldsymbol{\mathcal{F}}_t^{b_i}$, but its dimension is processed to d, e.g., $\boldsymbol{\mathcal{F}}_t^{o_1} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d}$. Moreover, pixel decoder also generates pixel embeddings $\boldsymbol{\mathcal{E}}_t^p \in \mathbb{R}^{H\times W\times d}$ of each frame. ❸ Subsequently, $\boldsymbol{\mathcal{F}}_t^{o_i}$ is used to attend cross attention of each transformer decoder layer in the query decoder for acquiring instance queries $\boldsymbol{\mathcal{Q}}_t \in \mathbb{R}^{N\times d}$ and masks $\boldsymbol{\mathcal{M}}_t \in \{0,1\}^{N\times H\times W}$, where N denotes the number of instances. The $\boldsymbol{\mathcal{Q}}_t$ is projected to instance embeddings $\boldsymbol{\mathcal{E}}_t \in \mathbb{R}^{N\times d}$ in CLIP space and calculates similarity with category embeddings $\boldsymbol{\mathcal{E}}^c \in \mathbb{R}^{C\times d}$ for recognizing instances.

Projector. To retain the semantic space of VLP model when projecting instance features, we refer to the design of SideAdapter (Xu et al. 2023). We mainly choose CLIP (Radford et al. 2021) as the VLP model, which is frozen all the time to avoid breaking its alignment space. At first, the pixel embeddings are projected to bias embeddings $\boldsymbol{\mathcal{E}}_t^b \in \mathbb{R}^{S\times H\times W\times d}$, where S denotes the number of attention heads. Subsequently, the instance queries $\boldsymbol{\mathcal{Q}}_t$ are used to dot product with the bias embeddings for getting attention biases $\boldsymbol{\mathcal{B}}_t \in \mathbb{R}^{N\times S\times H\times W}$. We copy the original class token of CLIP vision encoder and use the attention biases $\boldsymbol{\mathcal{B}}_t$ to guide the self-attention between the copied cls token and patch tokens in the last three CLIP layers. Finally, the copied cls token in CLIP vision output is the instance embeddings $\boldsymbol{\mathcal{E}}_t$. As for text projection, the category text set is directly projected by the CLIP text encoder to category embeddings $\boldsymbol{\mathcal{E}}^c$.

Resampler. The detailed calculation workflow of this simple temporal module is shown in Fig. 3, and we mainly refer to DVIS (Zhang et al. 2023b). The resampler is repeated L times for deep supervision. Before inputting Resampler, we match queries between adjacent frames via Hungarian algorithm (Kuhn 1955) for acquiring resorted video instance queries $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{N\times T\times d}$ and masks $\boldsymbol{\mathcal{M}} \in$

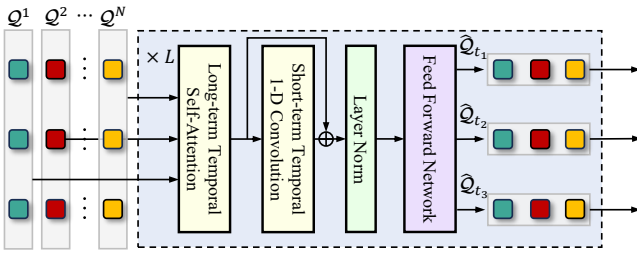

Figure 3: **Resampler.** The resampler serves as a lightweight parameterized temporal module to learn linking frame-level instance features as Brownian bridge. Resampler has L times repetition during calculation.

$\{0,1\}^{N\times T\times H\times W}$. Given resorted instance queries $\{\hat{\boldsymbol{\mathcal{Q}}}^1 \in \mathbb{R}^{T\times d}, \hat{\boldsymbol{\mathcal{Q}}}^2, \cdots, \hat{\boldsymbol{\mathcal{Q}}}^N\}$, we first input them into self-attention to make instance queries interact with each other globally. Then we input queries into 1-D convolution to aggregate local information of multiple adjacent frames, where the kernel size and stride of the convolution block are 5 and 1, respectively. To avoid vanishing the global property, we add a shortcut to replenish long-range context information from self-attention. The enhanced instance queries are normed by layernorm and are then processed by multiple linear layers.

Bridge-Text Alignment

BTA serves as a training mechanism and can be divided into two steps: (1) Linking instance across multiple frames as Brownian bridge; (2) Aligning instance Brownian bridge to class text. In the former step, we design **Head-Tail Matching** (Fig. 4a) for constraining bridge width between head and tail and design **Bridge Contrastive** (Fig. 4a) for molding independent frame-level instance features to follow the distribution of Brownian bridge. In the latter step, we design **Bridge-Text Contrastive** (Fig. 4b) for making instance Brownian bridge align with correct class label text.

Formally, given multiple sampled frames of a video $\boldsymbol{\mathcal{V}} = \{I_s, I_{s+1}, ..., I_e\}$, we construct data batch by randomly sampling one frame as triplets $\{I_s, I_t, I_e\}$, where $1 \le s < t < e \le T$. After the processing of the segmentor and resampler, we can acquire B batches instance queries $\hat{\boldsymbol{\mathcal{Q}}} \in \mathbb{R}^{B\times N\times 3\times d}$ and flatten the batch dimension to $\hat{\boldsymbol{\mathcal{Q}}} \in \mathbb{R}^{B\cdot N\times 3\times d}$. $\hat{\boldsymbol{\mathcal{Q}}}$ are then projected to instance embeddings $\hat{\boldsymbol{\mathcal{E}}} \in \mathbb{R}^{B\cdot N\times 3\times d}$. $B \cdot N$ is set as batch dimension.

Head-Tail Matching. Extract head $\hat{\boldsymbol{\mathcal{E}}}_s$ and tail $\hat{\boldsymbol{\mathcal{E}}}_e$ instance embeddings, we adopt a hinge loss to keep the distance between head and tail embeddings in a reasonable range:

$$\mathcal{L}_{htm} = \max\left(0, \Delta - \hat{\boldsymbol{\mathcal{E}}}_s \cdot \hat{\boldsymbol{\mathcal{E}}}_e\right), \tag{3}$$

$$= \mathrm{ReLU}\left(\Delta - \hat{\boldsymbol{\mathcal{E}}}_s \cdot \hat{\boldsymbol{\mathcal{E}}}_e\right), \tag{4}$$

where the embeddings is processed by ℓ_2 normalize, ReLU denotes rectified linear unit activation function, and Δ denotes the bound value. We convert ReLU to its smooth approximation (Glorot, Bordes, and Bengio 2011) to acquiring a gradient-friendly loss:

$$ReLU(z) = \max(0, z) \approx \log(1 + \mathrm{e}^z). \tag{5}$$

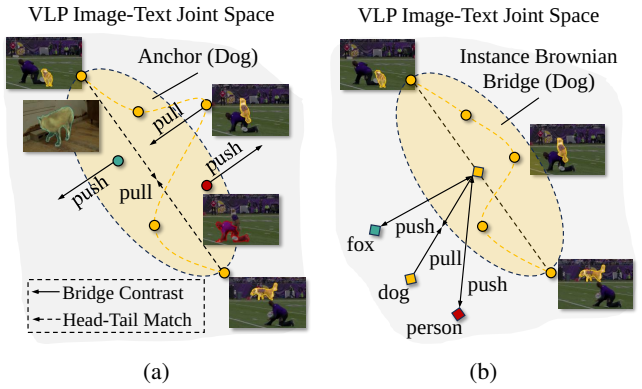

Figure 4: **Bridge-Text Alignment.** BTA serves as a training mechanism and can be divided into two steps: (1) Linking instances spanning multiple frames as a Brownian bridge; (2) Aligning instance Brownian bridge to class text. The first step is implemented via (a) Head-Tail Matching & Bridge Contrastive losses. The second step is achieved by (b) Bridge-Text Contrastive loss.

The final formula of matching loss is:

$$\mathcal{L}_{htm} = \log\left[1 + \mathrm{e}^{\Delta - \hat{\boldsymbol{\mathcal{E}}}_s \cdot \hat{\boldsymbol{\mathcal{E}}}_e}\right] \tag{6}$$

Bridge Contrastive. The goal of this objective is to make instance embeddings follow the Brownian bridge transition density in Eq. 1, which is ensured via a contrastive objective inspired by (Wang et al. 2022; Zhang et al. 2023a). Firstly, we define the contrastive distance measurement between middle states $\hat{\boldsymbol{\mathcal{E}}}_t$ and the target point in Brownian bridge:

$$d(\hat{\boldsymbol{\mathcal{E}}}_s^i, \hat{\boldsymbol{\mathcal{E}}}_t^i, \hat{\boldsymbol{\mathcal{E}}}_e^i) = -\frac{1}{2\sigma^2}\left\|\hat{\boldsymbol{\mathcal{E}}}_t^i - (1-\beta)\hat{\boldsymbol{\mathcal{E}}}_s^i - \beta\hat{\boldsymbol{\mathcal{E}}}_e^i\right\|_2^2, \tag{7}$$

where σ^2 is the variance in Eq. 1: $\frac{(t-s)(e-t)}{(e-s)}$, β is equal to $\frac{t-s}{e-s}$ and i is the instance index at batch axis, i.e. $i \in [0, B\cdot N - 1]$. Suppose that we set an instance embedding as anchor $\hat{\boldsymbol{\mathcal{E}}}^a$. To engrave the Brownian bridge of the anchor instance, the objective push away those negative middle states $\{\hat{\boldsymbol{\mathcal{E}}}_t^i | i \neq a\}$ which does not belong to anchor instance and pull the middle state of anchor instance to close the inner point of Brownian bridge. As described in previous works about contrastive learning (Robinson et al. 2021), we should pay more attention to pushing away those informative negative states so we introduce top-K strategy to selecting these confusing middle state as negative set $\boldsymbol{\mathcal{N}}$:

$$\boldsymbol{\mathcal{N}}^a = \{\hat{\boldsymbol{\mathcal{E}}}_t^j | j \in \mathrm{topK}(d(\hat{\boldsymbol{\mathcal{E}}}_s^a, \hat{\boldsymbol{\mathcal{E}}}_t^j, \hat{\boldsymbol{\mathcal{E}}}_e^a)), j \neq a\}, \tag{8}$$

$$\mathcal{L}_{bc}^a = -\log\frac{e^{d(\hat{\boldsymbol{\mathcal{E}}}_s^a, \hat{\boldsymbol{\mathcal{E}}}_t^a, \hat{\boldsymbol{\mathcal{E}}}_e^a)}}{e^{d(\hat{\boldsymbol{\mathcal{E}}}_s^a, \hat{\boldsymbol{\mathcal{E}}}_t^a, \hat{\boldsymbol{\mathcal{E}}}_e^a)} + \sum_{\hat{\boldsymbol{\mathcal{E}}}_t^j \in \boldsymbol{\mathcal{N}}^a} e^{d(\hat{\boldsymbol{\mathcal{E}}}_s^j, \hat{\boldsymbol{\mathcal{E}}}_t^j, \hat{\boldsymbol{\mathcal{E}}}_e^j)}}, \tag{9}$$

where K is regularly set to 5 in our experiments.

Bridge-Text Contrastive. To align the instance brownian bridge to class text embeddings $\boldsymbol{\mathcal{E}}^c$, according to the goal-conditioned nature, we represent instance via its Brownian

Method	Backbone	Vocabulary	BURST			LV-VIS	
			all	common	uncommon	val	test
Detic-O (Zhou et al. 2022)	R50	L-1203	2.7	2.8	1.8	7.7	7.0
OV2Seg (Wang et al. 2023)	R50	L-1203	3.7	3.9	2.4	14.2	11.4
OpenVIS (Guo et al. 2023)	R50	C-80,Y-40	3.5	5.8	3.0	12.0	10.2
BriVIS	R50	C-80,Y-40	**5.7** ↑ **2.0**	**12.9** ↑ **7.1**	**4.0** ↑ **1.0**	**20.9** ↑ **6.7**	**15.2** ↑ **3.8**
Detic-O (Zhou et al. 2022)	SwinB	L-1203	3.9	4.1	2.4	14.5	11.8
OV2Seg (Wang et al. 2023)	SwinB	L-1203	4.9	5.3	3.0	21.1	16.4
OpenVIS (Guo et al. 2023)	SwinB	C-80,Y-40	4.7	7.3	4.2	14.1	11.3
BriVIS	SwinB	C-80,Y-40	**7.0** ↑ **2.1**	**14.6** ↑ **7.3**	**5.1** ↑ **0.9**	**22.1** ↑ **1.0**	**16.7** ↑ **0.3**

Table 1: **Main results** on "Cross-dataset" evaluation protocol. **Bold** denotes the best performance. "Vocabulary" denotes the training dataset. "L-1203" denotes the LVIS (Gupta, Dollar, and Girshick 2019) dataset with 1203 categories. "Y-40" denotes the YouTube-VIS (Yang, Fan, and Xu 2019) dataset with 40 categories. "common" is a subset of the BURST (Athar et al. 2023) dataset consisting of 78 common object categories, while "uncommon" denotes another BURST subset composed of 404 uncommon object categories. "Detic-O" are Detic model (Zhou et al. 2022) with OWTB (Liu et al. 2022).

Method	Backbone	BURST			LV-VIS		
		base	novel	overall	base	novel	overall
OpenVIS (Guo et al. 2023)	R50	7.1	2.0	3.5	9.3	12.3	12.0
BriVIS	R50	**13.7** ↑ **6.6**	**2.8** ↑ **0.8**	**5.7** ↑ **2.2**	**22.2** ↑ **12.9**	**20.2** ↑ **7.9**	**20.9** ↑ **8.9**
OpenVIS (Guo et al. 2023)	SwinB	8.7	3.5	4.7	13.0	14.2	14.1
BriVIS	SwinB	**17.2** ↑ **8.5**	**3.7** ↑ **0.2**	**7.0** ↑ **2.3**	**22.3** ↑ **8.7**	**22.0** ↑ **7.8**	**22.1** ↑ **8.0**

Table 2: **Main results** on "Base-Novel" evaluation protocol. **Bold** denotes the best performance. The results of OpenVIS and BriVIS are corresponding to Tab. 1.

bridge center, i.e., the average of the head and tail instance embeddings $\hat{\boldsymbol{\mathcal{E}}}_{s,t,e} = (\hat{\boldsymbol{\mathcal{E}}}_s + \hat{\boldsymbol{\mathcal{E}}}_e)/2$. Then we adopt a contrastive objective to accomplish this alignment purpose:

$$\mathcal{L}_{btc} = -\log \frac{e^{\hat{\boldsymbol{\mathcal{E}}}_{s,t,e} \cdot \boldsymbol{\mathcal{E}}^c_{pos}}}{\sum_{k \in \boldsymbol{\mathcal{C}}} e^{\hat{\boldsymbol{\mathcal{E}}}_{s,t,e} \cdot \boldsymbol{\mathcal{E}}^c_k}}, \tag{10}$$

where $\boldsymbol{\mathcal{E}}^c_{pos}$ denotes the correct category of instance.

Overall Objective. Associating all of the losses above, we acquire the overall objective of BTA:

$$\mathcal{L}_{bta} = \frac{1}{B \cdot N} \sum_{a=0}^{B \cdot N} \boldsymbol{\mathcal{L}}^a_{htm} + \boldsymbol{\mathcal{L}}^a_{bc} + \boldsymbol{\mathcal{L}}^a_{btc}. \tag{11}$$

Discussion about BTA and other contrastive objectives. Previous contrastive objectives of VIS task try to pull the same instance across frames closer in embedding space (e.g., IDOL (Wu et al. 2022b), CTVIS (Ying et al. 2023)), which mainly aims at pursuing temporal consistency. However, this operation will gradually compress the instance features of different timestamps into one point. Although it can make VIS model track instances better, the dynamic information of instances is partially lost. The bridge contrastive of our BTA tries to align the instance feature and the feature point of the Brownian bridge at the same timestamp, which retains the temporal evolution of instance features for better capturing instance movement information. Therefore, our BTA can better process instances with rich temporal context than previous contrastive objectives. Moreover, since the middle instance features of the Brownian bridge are restricted between the start and the end state, previous contrastive objectives are a special case of our bridge contrastive when the distance between the start and end state is 0.

Experiments

Experiment Settings

Implementation Details. We select the mean Average Precision (mAP) from (Yang, Fan, and Xu 2019) as our evaluation metric. According to previous OVVIS works (e.g., OpenVIS (Guo et al. 2023)), we set Youtube-VIS 2019 and COCO (Lin et al. 2014) as our training dataset (101 categories). Following the Mask2former-VIS (Cheng et al. 2021), We resize the shorter side to either 360 or 480 and adopt a random horizontal flip strategy. To reduce training costs, we adopt two-stage training strategies. In the first stage, we randomly sample frames from departed videos and mix them with image data for training an open-vocabulary instance segmentor. The number of sampled frames T is set to 1. We train the open-vocabulary instance segmentation for 6k iterations with a batch size of 16 in this stage. Furthermore, we initial the model with the weights of Mask2Former (Cheng et al. 2022) pretrained on COCO (Lin et al. 2014) instance segmentation dataset. In the second stage, we train the whole model on coherent clips. To take

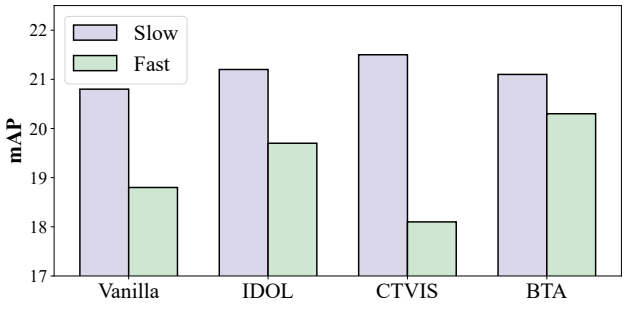

Figure 5: **Comparison** of OVVIS performance on slow and fast moving instances. Fast moving instances require model to have better temporal modeling ability. The overall mAP of them are 19.8, 20.1, 20.5, 20.9, respectively.

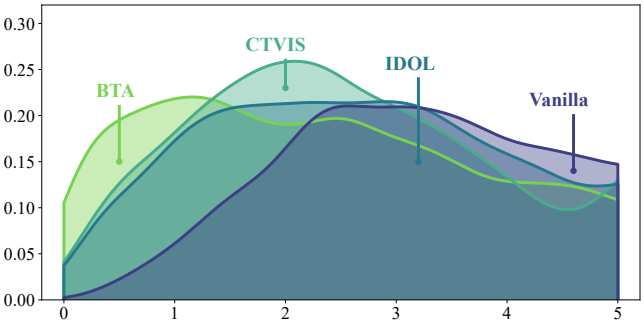

Figure 6: **Classification Entropy Distribution.** The entropy is calculated based on alignment scores between videos and class texts. If the learned representation has lower entropy, it denotes it is a more precise semantic descriptor.

full advantage of image data, we generate pseudo clips via deforming images and mix them with normal clips. We sample $T = 5$ frames from videos and train models for 6k iterations with a batch size of 16. AdamW (Loshchilov and Hutter 2019) is adopted as our optimizer, and the learning rate is set to 1e-4, which is scaled by a decay factor of 0.1 at the 5k iterations of two training stages. The number of queries N is set to 100. The bound value Δ is set to 0.5 by default. We adopt the "ViT-B/16"/"ViT-L/14" CLIP for those models with R50/Swin-Base backbone. When acquiring the class embedding, we input text in CLIP with prompt templates, e.g., "a photo of {category}" and ensemble 14 text prompt templates from ViLD (Gu et al. 2022).

Cross-Dataset Evaluation. We mainly refer to the challenge "Cross-Dataset" setting from image-level open-vocabulary segmentation (Xu et al. 2022). In this setting, the OVVIS model is trained on one dataset and evaluated on another dataset without any adapting. In this protocol, the mAP on all categories of BURST (482 categories) and LVVIS (1196 categories) are adopted as evaluation results.

Base-Novel Evaluation. In this protocol, we first split the categories of BURST and LVVIS into base part and novel part according to the CLIP similarity between class texts of evaluation and training dataset. BURST has 95 base categories and 387 novel categories. LVVIS has 124 base categories and 1072 novel categories. Afterward, we respectively calculate the mAP on base categories and novel categories as evaluation results.

Quantitative Analysis

Main Results. In Tab. 1, and Tab. 2 we compare our BriVIS to previous representative OVVIS methods. As for the "Cross-dataset" evaluation protocol, compared to previous OVVIS SOTA, our method exhibits 54.1% performance improvement on the challenge large-vocabulary BURST VIS dataset (5.7 vs. 3.7 mAP) and also achieves superior performance on LVVIS dataset, e.g., BriVIS achieves 6.7 mAP improvement compared to OV2Seg. As for the "Base-novel" evaluation protocol, compared to OpenVIS, we obtain obviously better performance, especially for the novel part of LVVIS where we improve OpenVIS from 12.3 to 20.2 mAP. Furthermore, on the larger visual backbone (e.g., Swin-Base), our method is still capable of surprising improvements, which further demonstrates the superior open-vocabulary ability of our method.

Frames	$\mathcal{L}_{bta}$	BURST all	LV-VIS val	Speed (s/iter)
all		4.4	19.8	2.77
all	✔	**5.7** ↑**1.3**	**21.1** ↑**1.3**	
head-tail		3.8	17.0	**1.88** ↓**0.89**
head-tail	✔	**5.7** ↑**1.9**	20.9 ↑**3.9**	

Table 3: **Ablation Study** of our core designs (BTA). "Frames" denotes the frame extraction scheme. "all" denotes using all of the video frames for aligning with texts. "head-tail" denotes classifying the video instance by only the first frame and the last frame when this instance appears.

Ablation Study. To check the effectiveness of our proposed BTA, we ablate this module from BriVIS in Tab. 3. The BTA can guide instance queries spanning video to follow the Brownian bridge for enhancing temporal learning, which improves baselines by 1.3 mAP on LVVIS and 1.3 mAP on BURST. Moreover, to verify the significance of BTA for reducing inference cost, we attempt a low-cost inference scheme, i.e., "head-tail" frame settings, which denotes classifying the video instance by only the first frame and the last frame when this instance appears. Training w/o $\mathcal{L}_{bta}$ will sharply decrease by 2.8 mAP compared to "all" frame settings (19.8 vs 17.0). However, training w/ $\mathcal{L}_{bta}$ only decrease 0.2 mAP (21.1 vs 20.9) and reduce 32% cost (2.77 $\rightarrow$ 1.88 s/iter), which demonstrates the significance.

Comparison between BTA and other contrastive methods. We first evaluate the overall performance of BTA and other contrastive methods on our baseline. As shown in Fig. 5, our BTA achieves the best open-vocabulary performance, which shows our method is more effective for OVVIS. Afterward, we analyze the semantic precision of different alignment methods via classification entropy of alignment score. As shown in Fig. 6, we plot the clas-

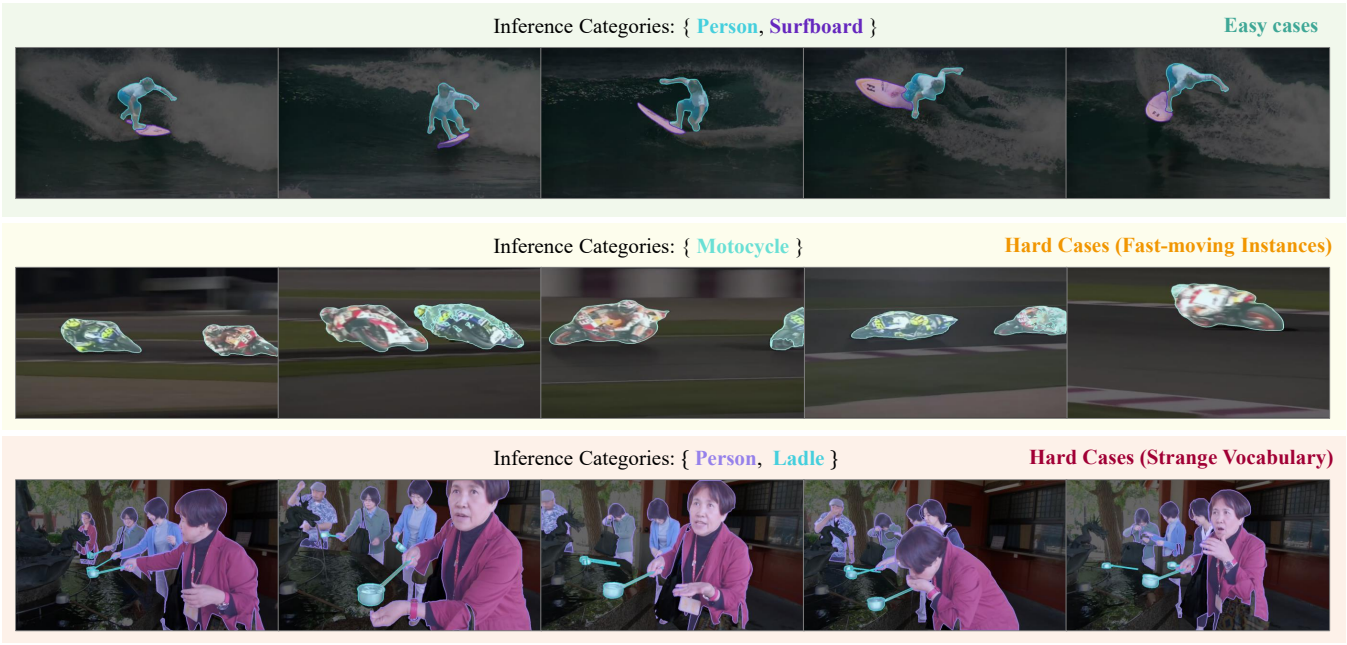

Figure 7: **Qualitative results of different cases.** We select easy cases in normal scenarios, hard cases in fast-moving instance video scenarios, and hard cases in strange vocabulary scenarios to verify the effectiveness of ours.

sification entropy distribution of Baseline (vanilla training), IDOL, CTVIS, and BriVIS. BTA provides lower entropy than CTVIS, IDOL, and baseline, which means the learning mode of BTA can better help OVVIS model extract discriminative features. Finally, to evaluate the temporal modeling capabilities, we categorize instances into fast-moving instances with rich temporal information and slow-moving instances with almost no temporal information by statistically measuring the distance moved by the bounding boxes. As shown in Fig. 5, BTA delivers mediocre performance on slow-moving instances and superior performance on fast-moving instances, which proves that our method has stronger temporal modeling capabilities and is consistent with the conclusion of discussions in Section Bridge-Text Alignment. According to three experiments above, we conclude that BTA is a contrastive objective that enables OVVIS model to learn more informative patterns, especially for temporal patterns. As a result, it allows the OVVIS model to better handle instances with rich temporal information.

Qualitative Analysis

As described in Section Quantitative Analysis, the method with bridge-text alignment (BriVIS) is more capable of processing fast-moving video instances (Fig. 5) and provides more robust alignment between video instances and class texts (Fig. 6) than baseline. To qualitatively verify this point, we select some easy and hard cases to illustrate the temporal segmentation performance. The green part of Fig. 7 shows that our method generates high-quality spatial-temporal masks under regular easy cases. The yellow part of Fig. 7 shows that our method can precisely process fast-moving video instances, which justifies the temporal modeling ability of our BTA. The red part of Fig. 7 shows that our method can recognize long-tail categories, which justifies the robust and precise alignment between bridge representation and text representation.

Conclusion

In this paper, we propose BriVIS to support considering instance movement context when using the image-text pretraining model to perform open-vocabulary recognition. Specifically, the BriVIS links independent frame-level instance features as Brownian bridge and aligns the bridge with text in image-text pretraining space for classifying instances at bridge level. The key of our method is to sublimate the granularity of instance descriptor from frame to bridge. Bridge-level instance representation is able to carry more context information, especially for instance movement context. Therefore, our method not only can better process fast-moving instances which require better temporal modeling, but also exhibits a more robust and precise alignment between instance and class texts. In addition, because of the goal-conditioned property of Brownian bridge modeling, our method only needs to infer a few frames to achieve the comparable performance as inferring all frames.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201), the Shenzhen Medical Research Funds in China (No. B2302037), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465, 6240071660, U24B600013), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

References

Athar, A.; Luiten, J.; Voigtlaender, P.; Khurana, T.; Dave, A.; Leibe, B.; and Ramanan, D. 2023. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1674–1683.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229. Springer.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4964–4973.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 540–557. Springer.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323. JMLR Workshop and Conference Proceedings.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *Proceedings of the International Conference on Learning Representations*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*.

Guo, P.; Huang, T.; He, P.; Liu, X.; Xiao, T.; Chen, Z.; and Zhang, W. 2023. OpenVIS: Open-vocabulary Video Instance Segmentation. *arXiv preprint arXiv:2305.16835*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5356–5364.

Han, C.; Zhong, Y.; Li, D.; Han, K.; and Ma, L. 2023. Open-Vocabulary Semantic Segmentation with Decoupled One-Pass Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1086–1096.

Han, S. H.; Hwang, S.; Oh, S. W.; Park, Y.; Kim, H.; Kim, M.-J.; and Kim, S. J. 2022. VISOLO: Grid-Based Space-Time Aggregation for Efficient Online Video Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2896–2905.

Heo, M.; Hwang, S.; Hyun, J.; Kim, H.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2023. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14623–14632.

Heo, M.; Hwang, S.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2022. VITA: Video Instance Segmentation via Object Token Association.

Huang, D.-A.; Yu, Z.; and Anandkumar, A. 2022. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*.

Hui, T.; Liu, S.; Ding, Z.; Huang, S.; Li, G.; Wang, W.; Liu, L.; and Han, J. 2023. Language-Aware Spatial-Temporal Collaboration for Referring Video Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huynh, D.; Kuen, J.; Lin, Z.; Gu, J.; and Elhamifar, E. 2022. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7031.

Hwang, S.; Heo, M.; Oh, S. W.; and Kim, S. J. 2021. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34: 13352–13363.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, M.; Li, S.; Xiang, W.; and Zhang, L. 2023. MDQE: Mining Discriminative Query Embeddings to Segment Occluded Instances on Challenging Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10524–10533.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Liu, Y.; Zulfikar, I. E.; Luiten, J.; Dave, A.; Ramanan, D.; Leibe, B.; Ošep, A.; and Leal-Taixé, L. 2022. Opening up Open World Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19045–19055.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Revuz, D.; and Yor, M. 2013. *Continuous martingales and Brownian motion*, volume 293. Berlin: Springer Science & Business Media.

Robinson, J. D.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2021. Contrastive Learning with Hard Negative Samples. In *International Conference on Learning Representations*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Tu, Y.; Zhang, X.; Liu, B.; and Yan, C. 2017. Video description with spatial-temporal attention. In *Proceedings of the 25th ACM international conference on Multimedia*, 1014–1022.

Wang, H.; Wang, S.; Yan, C.; Jiang, X.; Tang, X.; Hu, Y.; Xie, W.; and Gavves, E. 2023. Towards Open-Vocabulary Video Instance Segmentation. *arXiv preprint arXiv:2304.01715*.

Wang, J.; Lin, D.; and Li, W. 2023. Dialogue Planning via Brownian Bridge Stochastic Process for Goal-directed Proactive Dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, 370–387. Toronto, Canada: Association for Computational Linguistics.

Wang, R. E.; Durmus, E.; Goodman, N.; and Hashimoto, T. 2022. Language modeling via stochastic processes. In *International Conference on Learning Representations*.

Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8741–8750.

Wu, J.; Jiang, Y.; Bai, S.; Zhang, W.; and Bai, X. 2022a. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, 553–569. Springer.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022b. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, 588–605. Springer.

Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; and Akata, Z. 2019. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8256–8265.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5188–5197.

Ying, K.; Zhong, Q.; Mao, W.; Wang, Z.; Chen, H.; Wu, L. Y.; Liu, Y.; Fan, C.; Zhuge, Y.; and Shen, C. 2023. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 899–908.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhang, H.; Liu, D.; Zheng, Q.; and Su, B. 2023a. Modeling Video As Stochastic Processes for Fine-Grained Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2225–2234.

Zhang, T.; Tian, X.; Wu, Y.; Ji, S.; Wang, X.; Zhang, Y.; and Wan, P. 2023b. DVIS: Decoupled Video Instance Segmentation Framework. *arXiv preprint arXiv:2306.03413*.

Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; and Torralba, A. 2017. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2002–2010.

Zheng Ding, Z. T., Jieke Wang. 2023. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In *International Conference on Machine Learning*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting Twenty-thousand Classes using Image-level Supervision. In *Proceedings of the European Conference on Computer Vision*.