Sliced Wasserstein Bridge for Open-Vocabulary Video Instance Segmentation

Anonymous ICCV submission

Paper ID 2556

Abstract

In recent years, researchers have explored the task of open-vocabulary video instance segmentation, which aims to identify, track, and segment any instance within an open set of categories. The core challenge of Open-Vocabulary VIS lies in solving the cross-domain alignment problem, including spatial-temporal and text-visual domain alignments. Existing methods have made progress but still face shortcomings in addressing these alignments, especially due to data heterogeneity. Inspired by metric learning, we propose an innovative Sliced Wasserstein Bridging Learning Framework. This framework utilizes the Sliced Wasserstein distance as the core tool for metric learning, effectively bridging the four domains involved in the task. Our innovations are threefold: (1) Domain Alignment: By mapping features from different domains into a unified metric space, our method maintains temporal consistency and learns intrinsic consistent features between modalities, improving the fusion of text and visual information. (2) Weighting Mechanism: We introduce an importance weighting mechanism to enhance the discriminative ability of our method when dealing with imbalanced or significantly different data. (3) High Efficiency: Our method inherits the computational efficiency of the Sliced Wasserstein distance, allowing for online processing of large-scale video data while maintaining segmentation accuracy. Through extensive experimental evaluations, we have validated the robustness of our concept and the effectiveness of our framework.

1. Introduction

In recent years, video instance segmentation methods [25–27] have demonstrated significant improvements and breakthroughs in various application fields. However, most existing methods rely on the **Closed-Set Assumption** [4], where the learning and recognition scope of the model are strictly limited to known categories or labels during training and testing. This assumption ignores the possibility of unknown categories in the real world, limiting the model's ability to effectively detect unknown categories. To overcome this limitation, researchers have begun to explore the task of open-vocabulary video instance segmentation [15, 24] (Open-Vocabulary VIS), which aims to identify, track, and segment any instance within an open set of categories. These Open-Vocabulary VIS methods can not only handle traditional predefined categories but also flexibly adapt to new user-specified categories, achieving the advanced function of on-demand segmentation.

The core of Open-Vocabulary VIS methods lies in solving the cross-domain alignment problem, including spatial-temporal domain alignment and text-visual domain alignment. Existing methods have different focuses. In terms of spatial-temporal domain alignment, current methods extract class-agnostic instance features based on the foundational model Mask2Former-VIS. Furthermore, OV2Seg [24], OpenVIS [15], and CLIP-VIS [31] employ memory-based approaches to perform cross-frame instance matching. In terms of text-visual domain alignment, existing methods mainly rely on the zero-shot performance exhibited by visual-language models, i.e., CLIP [22] pretrained on large-scale image-text pairs. Specifically, BriVIS [8] and OV-Former [11] introduce additional CLIP image encoders on top of visual feature extractors. The difference lies in that BriVIS integrates them based on contrastive learning, while OV-Former maps text and visual features into the same space vis the attention mechanism.

Despite the significant progress made by existing Open-Vocabulary VIS methods, there are still many shortcomings. The cross-domain alignment problem remains inadequately addressed, and the inherent data heterogeneity results in significant differences in feature distributions across different domains. In light of this, we cannot help but trace back to the development history of traditional machine learning methods, seeking a tool that can simultaneously solve spatial-temporal domain alignment and text-visual domain alignment. We find that the metric learning method in the field of machine learning provides an effective approach to addressing the alignment problem [12, 17]. By learning a suitable metric space, this paradigm makes similar samples closer while dissimilar samples relatively farther apart, providing a powerful tool for solving the alignment problem.

Guided by the idea of metric learning, we propose an innovative Sliced Wasserstein Bridging Learning Frame-

work, i.e., SWbridge. This framework utilizes the Sliced Wasserstein distance [3] as the core tool for metric learning. This framework cleverly utilizes the Sliced Wasserstein distance as the core tool for metric learning. Specifically, we minimize the Sliced Wasserstein distance between samples based on random paths through sampling, achieving feature embedding alignment. We bridge four domains based on a strategy to address the challenges of cross-domain alignment (spatial-temporal domain and text-visual domain).

Our innovations are mainly concentrated in three aspects: **(1) Domain Alignment.** The Sliced Wasserstein distance has the advantage of accurately capturing differences in data distributions in high-dimensional spaces. Our method uses the Sliced Wasserstein distance as a bridge to map features from different domains into a unified metric space. This not only maintains temporal consistency by comparing instance embeddings between adjacent frames, achieving temporal alignment and effectively mitigating challenges such as progressive occlusion, but also learns intrinsic consistent features between modalities, narrowing the semantic gap and improving the fusion of text and visual information through modality alignment. **(2) Weighting Mechanism:** We introduce an importance weighting mechanism to further enhance its discriminative ability when dealing with imbalanced data or data with significant feature differences. In open-vocabulary video instance segmentation tasks, there may be significant feature differences between different instances, and some instances may be more critical or informative. Through the importance weighting mechanism, our method can dynamically adjust the contribution of different instances in the metric space, allowing key instances to have greater weight in distance calculations. **(3) High Efficiency:** Our SWbridge inherits the computational efficiency advantage of the Sliced Wasserstein distance, making it highly efficient when processing large-scale video data. Without introducing additional network parameters, our method can handle video in an online manner while maintaining segmentation accuracy.

Combining the aforementioned perspectives and innovations, we propose an open-vocabulary video instance segmentation method capable of bridging multiple domains. This method has achieved remarkable results on multiple datasets related to open-vocabulary and video tasks, such as LV-VIS [24], YT-VIS 2019/2021 [26], BURST [1], and OVIS [21]. Through a series of comprehensive ablation studies in §5.4, our extensive experimental evaluations have fully validated the robustness of our concept and the effectiveness of our framework.

2. Related Works

2.1. Open-Vocabulary Detection and Segmentation

Open Vocabulary Detection (OVD) [9, 23] and Open Vocabulary Segmentation (OVS) [13, 18] are cutting-edge tasks in the field of computer vision, enabling models to be trained on images containing unannotated novel objects, thereby breaking the closed-set constraint. This breakthrough is primarily attributed to the application of weak supervision signals, namely the utilization of image-text pairs (such as image-caption pairs and image-level labels) or large pre-trained Visual-Language Models (VLMs), such as CLIP [22]. Supported by weak supervision signals, methods for OVD and OVS can be categorized into four main types [30]. The first type is region-aware training, which learns object feature representations by exploring the intrinsic links between image regions and text descriptions without relying on the VLMs' image encoder (VLMs-IE) or direct object annotations. The second type is the pseudo-labeling method, which enhances model generalization by generating pseudo-labels from preliminary predictions on unannotated images using both image-text pairs and the VLMs-IE. The other two types are knowledge distillation and transfer learning. Knowledge distillation transfers knowledge from pre-trained VLMs to new models or tasks via a distillation mechanism, while transfer learning applies knowledge learned by pre-trained models on specific tasks to new tasks or domains. Both rely on the VLMs-IE and seldom involve direct training on image-text pairs.

Furthermore, zero-shot learning [28] addresses the closed-set constraint in scene-aware tasks and differs from open vocabulary methods. It prohibits access to weak supervision signals during training, but can be converted into Open-Vocabulary tasks upon gaining access, enhancing the model's generalization and recognition of novel objects.

2.2. Open-Vocabulary Video Instance Segmentation

Open-vocabulary video instance segmentation, as an emerging vision task, aims to simultaneously classify, track, and segment arbitrary objects within open categories in videos, attracting considerable attention from researchers in recent years. OV2Seg [24] has taken the lead in this field by not only constructing the Large Vocabulary Video Instance Segmentation (LV-VIS) dataset but also proposing the first end-to-end benchmark for open-vocabulary video instance segmentation. OV2Seg method leverages the Mask2Former [6] framework to extract class-agnostic masks and query embeddings, utilizes the CLIP text encoder for precise mask classification, and achieves cross-frame instance tracking through a long-term matching strategy. Subsequent research has mostly followed this basic approach and expanded upon it. OpenVIS [15] introduces a two-stage framework, OVIS, which adopts a neighboring matching strategy for instance tracking, effectively simplifying the processing pipeline. BriVIS [8] models instances as Brownian bridges and closely aligns bridge-level instance representations with category texts through contrastive learning, also employing a neighboring matching strategy to enhance the accuracy of instance recognition. OVFormer [11]

introduces a novel unified embedding alignment module that effectively addresses the domain gap between instance queries and text embeddings, demonstrating good practicality with its semi-online processing approach. However, the aforementioned methods primarily focus on bridging the domain discrepancy between text and vision, with insufficient attention to spatiotemporal consistency among cross-frame instances. CLIP-VIS [31] improves upon this by introducing a time-topK enhanced matching strategy, which strengthens temporal modeling capabilities between frames and improves matching accuracy.

We believe that the core of the open-vocabulary video instance segmentation task lies in simultaneously addressing the alignment of the spatial-temporal domain and the text-visual domain. To this end, based on the concept of metric learning, we employ the Sliced Wasserstein distance to bridge various domains. This approach not only effectively narrows the domain gap between modalities but also ensures spatio-temporal consistency of instances across frames, providing a novel solution for the open-vocabulary video instance segmentation task.

2.3. Vision-Language Models

The core of Visual Language Models (VLMs) [5] is their deep training on large-scale image-text paired data, enabling the fusion and understanding of visual and textual information, and endowing them with powerful zero-shot object recognition capabilities [19]. CLIP [22], a prominent VLM, uses an image encoder to extract visual features and a text encoder to generate text embeddings, jointly constructing a cross-modal representation space. However, applying VLMs to open-vocabulary video instance segmentation is challenging. VLMs, primarily trained on images, struggle with understanding dynamic video scenes rich in spatiotemporal information, and lack the capability to maintain object consistency across frames. Additionally, the domain shift between text and vision poses a significant challenge, potentially leading to biases in mapping textual information to the visual space. To overcome these limitations, we propose a progressive approach that addresses the spatiotemporal consistency issue and the domain shift problem, enabling VLMs to better adapt to video instance segmentation tasks.

3. Preliminaries

3.1. Open-Vocabulary VIS Formulation

Given a test video $\mathcal{D}_{test}$ with T frames, the objective in Open-Vocabulary VIS is to accurately predict all N instances belonging to the categories in $\mathcal{C} = \mathcal{C}_{base} \cup \mathcal{C}_{novel}$ by the trained model f_θ. $\mathcal{C}_{base}$ is the set of base (training) categories, and $\mathcal{C}_{novel}$ is the set of novel categories that are not seen during training but may appear in the test videos. The prediction can be formulated as:

$$\{\{m_1, m_2, \ldots, m_T\}, c\}_n^N = f_\theta(\mathcal{D}_{test}), \tag{1}$$

where $\{m_t\}_{t=1}^T$ is the segmentation masks, and $c \in \mathcal{C}$.

3.2. Sliced Wasserstein Distance

One-dimensional Wasserstein Distance. For one-dimensional probability measures μ and ν in $\mathcal{P}_p(\mathbb{R})$, the p-Wasserstein distance is defined as:

$$\mathrm{W}_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p \, dz, \tag{2}$$

where F_μ and F_ν are the cumulative distribution functions (CDFs) of μ and ν. This formulation provides a closed form for computing the Wasserstein distance in one-dimensional spaces, making it well-suited for projected measures.

Sliced Wasserstein Distance. To generalize the Wasserstein distance to higher-dimensional measures, the Sliced Wasserstein Distance (SWD) projects the measures μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ onto one-dimensional subspaces, and then averages the one-dimensional Wasserstein distances from these projections. For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the SWD is defined as:

$$\mathrm{SW}_p^p(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})} \left[\mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \right], \tag{3}$$

where $\theta\sharp\mu$ and $\theta\sharp\nu$ represent the push-forward measures of μ and ν along direction $\theta \in \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. This projection $f(x) = \theta^\top x$ maps points from $\mathbb{R}^d$ to $\mathbb{R}$, enabling the computation of Wasserstein distances in a one-dimensional space.

Since the expectation in Eq. 3 is computationally expensive, SWD is typically approximated by averaging over L independent directions, $\theta_1, \ldots, \theta_L$, sampled from $\mathcal{U}(\mathbb{S}^{d-1})$:

$$\widehat{\mathrm{SW}}_p^p(\mu, \nu; L) = \frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu), \tag{4}$$

where each $\theta_l\sharp\mu$ and $\theta_l\sharp\nu$ are projected representations of μ and ν along the direction θ_l. The number of projections L controls the accuracy of the Monte Carlo approximation.

4. Method

As illustrated in Fig. 2, we propose the SWbridge framework. Firstly, in §4.1, we integrate a category-agnostic feature extractor. Then, in §4.2, we introduce a spatio-temporal bridging module that robustly establishes instance associations between adjacent frames, ensuring the consistency and coherence of temporal information. Finally, in §4.3, we design a modal bridging module that performs fine-grained semantic mapping across the text and visual domains to enhance cross-domain understanding and effectively address the challenges posed by domain shifts.

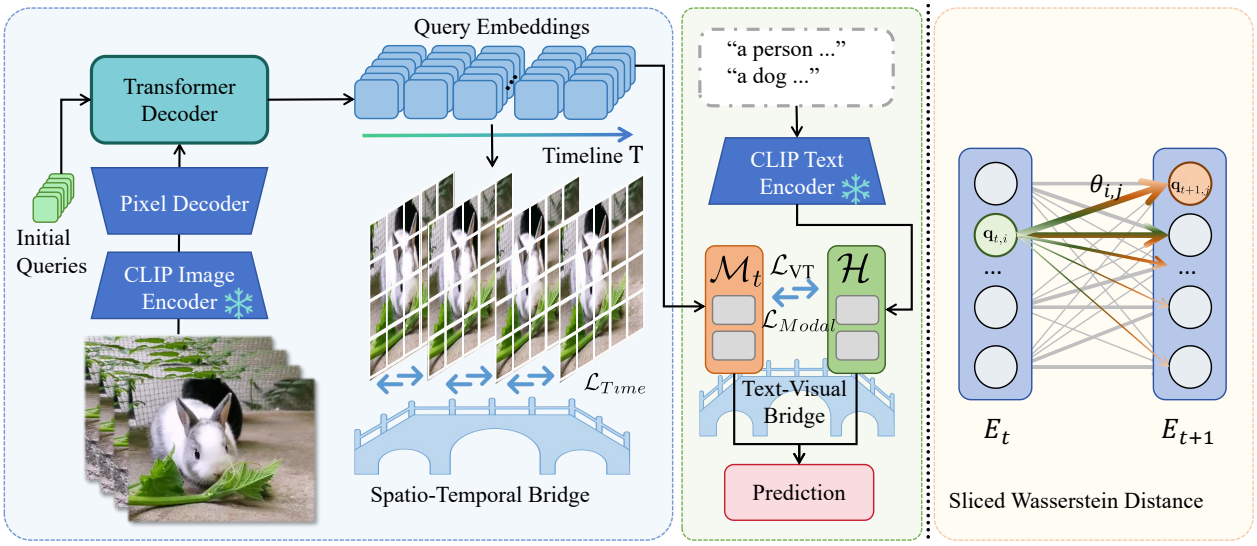

Figure 1. **Overview of the proposed SWBridge framework (left) and the Sliced Wasserstein distance calculation process (right).**

4.1. Class-agnostic Feature Extractor

Given a video $\mathcal{D}$ consisting of T frames, $\{V_t \in \mathbb{R}^{3\times H\times W}\}_{t=1}^{T}$, where H and W denote the height and width of each frame, our model follows the conventional paradigm [11, 24, 31] by adopting the Mask2Former architecture as a class-agnostic feature extractor. The Mask2Former comprises three core components: an image encoder f_{CLIP}, a pixel decoder f_{decode}, and a Transformer decoder $f_{\text{Transformer}}$. Specifically, the frozen CLIP model f_{CLIP} processes each frame V_t independently to extract high-level visual features F_t. These features are then input to the pixel decoder f_{decode}, which upsamples them to produce a multi-scale feature map D_t preserving spatial details. Furthermore, a set of learnable queries Q is initialized and fed into the Transformer decoder $f_{\text{Transformer}}$ along with the high-resolution feature map D_t. The Transformer decoder processes these inputs to generate the query embeddings E_t.

4.2. Spatio-Temporal Bridge

After extracting category-agnostic features, the next challenge is to construct a spatio-temporal bridge to establish the correspondence between instances in consecutive frames t and $t+1$. Given the dynamic nature of video content, objects may encounter occlusion, movement, and appearance changes. Therefore, it is particularly important to ensure the accuracy of instance tracking over time. To achieve this, we employ the Sliced Wasserstein metric.

Initially, for the query embeddings of frames t and $t+1$, we perform random sampling to construct random paths $Z_{i,j} = \mathbf{q}_{t,i} - \mathbf{q}_{t+1,j}$, where $\mathbf{q}_{t,i} \in E_t$ and $\mathbf{q}_{t+1,i} \in E_{t+1}$ are outputs of the Transformer decoder $f_{\text{Transformer}}$. These paths are then normalized to obtain the unit direction $\theta_{i,j} = \frac{Z_{i,j}}{\|Z_{i,j}\|_2}$, which serves as the projection direction.

Subsequently, we sample a new projection direction $\hat{\theta}_{i,j}$ from $\sigma_\kappa(\theta_{i,j}; \frac{Z_{i,j}}{\|Z_{i,j}\|_2})$. This introduces randomness, potentially yielding different projection directions each time, which aids in exploring multiple relative positional relationships between feature points. This, in turn, may enhance the model's robustness and generalization capability.

With the projection direction $\hat{\theta}_{i,j}$ in hand, we proceed to calculate the one-dimensional Wasserstein distance. This involves projecting $\mathcal{M}_t$ and $\mathcal{T}$ along the θ direction to obtain one-dimensional distributions $\theta\sharp\mu$ and $\theta\sharp\nu$. These distributions are then sorted, and the inverse functions of their cumulative distribution functions (CDFs) are computed. Finally, the one-dimensional Wasserstein distance $W_p^p(\theta\sharp\mu, \theta\sharp\nu)$ is calculated using the formula:

$$W_p^p(F, G) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p \, dt, \tag{5}$$

where F and G are the CDFs of $\theta\sharp\mu$ and $\theta\sharp\nu$, respectively, and p is typically set to 2.

The next step involves importance weighting and Monte Carlo estimation. We repeat the aforementioned process L times to obtain L projection directions $\{\theta_1, \ldots, \theta_L\}$ and the corresponding set of one-dimensional Wasserstein distances. Each distance is then weighted with w, where w is determined by an increasing function $f(\cdot)$. The weighted average of these distances serves as the approximation of SW, calculated using the formula:

$$SW = \frac{1}{L}\sum_{l=1}^{L} w \cdot \frac{f(W_p^p(\theta_l\sharp\mathcal{M}_t, \theta_l\sharp\mathcal{T}))}{\sum_{j=1}^{L} f(W_p^p(\theta_j\sharp\mathcal{M}_t, \theta_j\sharp\mathcal{T}))}, \tag{6}$$

where the weight w for each projection direction is:

$$w = \frac{f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{j=1}^{L} f(W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu))}. \tag{7}$$

Eq.6 yields a weighted Sliced Wasserstein distance that reflects the relationship between adjacent frames along random projection directions, considering both the geometric positions of data points and their differences through importance weighting.

To establish instance correspondences, we compute the SW values between all pairs of instances across consecutive frames t and $t+1$, denoted as $\mathrm{SW}_p^p(E_t, E_{t+1})$. These values quantify the difference between instances along the projection direction, with smaller SW values indicating stronger spatio-temporal consistency. By selecting the instance pairs with the smallest SW values, we identify those with the highest temporal alignment, ensuring robust and meaningful associations for instance tracking. Formally, the instance matching objective is:

$$\mathcal{L}_{Time} = \arg\min \ \mathrm{SW}(E_t, E_{t+1}), \tag{8}$$

where the selected pairs correspond to instances with the strongest spatio-temporal consistency, facilitating accurate tracking and correlation across frames. Our method thus emphasizes directions and features that capture significant inter-frame changes, effectively prioritizing instances that reflect spatial and appearance consistency. The adaptive weighting in Eq.6 further enhances tracking by highlighting subtle variations, ensuring robust temporal consistency.

4.3. Text-Visual Bridge

After bridging the temporal consistency between the spatial and temporal domains through the spatio-temporal bridge, we are still confronted with the challenge of addressing the modal domain shift between the text and visual domains. Video content, with its dynamic and diverse nature, introduces additional complexity as textual descriptions and visual appearances can vary significantly. To ensure the accurate alignment between text and visual features, which is crucial in this context, we continue to leverage the Sliced Wasserstein metric. This metric has proven effective in our previous spatio-temporal bridging efforts, leading us to adopt it for modal alignment in our current method.

Specifically, we first combine the temporal feature D_t with the query embeddings E_t through multiplication to obtain the visual embeddings M_t. Subsequently, we apply a mask pooling operation to derive a class-agnostic mask $\mathcal{M}_t \in \mathbb{R}^{N \times D}$ based on M_t and the frame features F_t. Lastly, we obtain the class text embeddings $\mathcal{H} \in \mathbb{R}^{L \times D}$ generated by the CLIP text encoder, where L represents the number of categories.

It is worth noting that our cross-modal alignment method bears similarity to the Sliced Wasserstein metric procedure detailed in §4.2. We begin by defining the projection path of the cross-modal module as $Z_{i,j}^c = \boldsymbol{m}_{t,i} - \boldsymbol{h}_{t,j}$. To obtain the projection direction, we normalize this path, resulting in $\theta_{i,j}$. Given the relatively small sizes of $\mathcal{M}_t$ and $\mathcal{H}$, this module departs from the random sampling approach and instead adopts a full-path full-projection strategy. Upon successfully obtaining the projection direction θ, we proceed to calculate the one-dimensional Wasserstein distance. We then apply importance weighting and incorporate the Monte Carlo estimation method in our subsequent calculations. Following the logic of Eq. 6 and Eq. 7, we ultimately compute the weighted average as an approximation of the Sliced Wasserstein distance.

Additionally, similar to Eq. 8, we introduce an additional loss function $\mathcal{L}_{Modal}$ to constrain the modality alignment process. This loss function is defined as:

$$\mathcal{L}_{Modal} = \arg\min \ \mathrm{SW}(\mathcal{M}_t, \mathcal{H}). \tag{9}$$

4.4. Loss Function

To mine similarity relations in cross-text video pairs, we draw on the successful experience of cross-modal learning, as demonstrated in [8]. To this end, we introduce a cross-text similarity alignment loss based on contrastive learning. The loss function is formulated as follows:

$$\mathcal{L}_{\mathrm{VT}} = -\frac{1}{NL} \sum_{i,j} \log \frac{\exp(d_{ij}/\tau)}{\sum_k \exp(d_{ik}/\tau) + \sum_l \exp(d_{jl}/\tau)}, \tag{10}$$

where $d_{ij} = \mathrm{sim}(\boldsymbol{m}_i, \boldsymbol{h}_j)$, $\mathrm{sim}(\cdot,\cdot)$ represents the similarity metric function, and τ is the temperature hyperparameter.

The **Total Loss** $\mathcal{L}_{\mathrm{Total}}$ is a weighted combination of multiple loss components, which can be written as:

$$\mathcal{L}_{\mathrm{Total}} = \alpha_T \mathcal{L}_{\mathrm{Time}} + \alpha_M \mathcal{L}_{\mathrm{Modal}} + \alpha_V \mathcal{L}_{\mathrm{VT}} + \alpha_S \mathcal{L}_{\mathrm{Seg}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{Seg}}$ represents the binary cross-entropy and dice losses for mask prediction, following the approach in [24]. The coefficients α_T, α_M, α_V, and α_S are the corresponding weighting factors for each loss component.

Computational complexity. In our Sliced Wasserstein metric process, the core steps involve matrix operations and sorting procedures. Initially, we generate a set of initial projection vectors by randomly selecting samples and computing the differences among these samples. The complexity of this selection operation is $O(L)$, while the complexity of the difference calculation is $O(L \cdot D)$, where L represents the number of projection vectors and D is the dimensionality of the samples. Subsequently, we proceed to the normalization and resampling phase, which also maintains a complexity of $O(L \cdot D)$. When calculating the one-dimensional Wasserstein distance, we need to perform matrix multiplication, a step with a complexity of $O(N \cdot D \cdot L)$, where N is the number of samples. Additionally, we need to sort the projected data, with a sorting operation complexity of $O(N \log N)$.

Table 1. **Statistics of the dataset.** Cat. and Ins. represent the number of categories and instances respectively.

Dataset	Videos	Cat.	Ins.	Train	Val.	Test
LV-VIS	4,828	1,196	-	3,083	837	908
YT-VIS 19	2,883	40	4,883	2,238	302	343
YT-VIS 21	3,859	40	8,171	2,985	421	453
OVIS	901	25	5,223	607	140	154
BURST	2,914	482	16,000	500	993	1,421

Although in practical applications, since L is usually much smaller than N, the impact of the sorting operation on the overall complexity may be relatively minor. In summary, the overall complexity of our method can be approximated as $O(N \cdot D \cdot L + N \log N)$. In practical applications, as the number of samples N and the number of projection vectors L increase, the computational cost of matrix multiplication will become dominant. Although sorting and other operations with linear complexity also exist, their contribution to the overall complexity is relatively small.

5. Experiments

5.1. Datasets and Evaluation Metrics

We initially train our method on a combined set of common and frequent categories derived from the LVIS dataset [16]. Subsequently, we assess the performance of our method on both the validation and test sets of the LV-VIS dataset [24], as well as the validation set of several datasets, namely OVIS [21], YT-VIS 2019/2021 [26], and BURST [1]. The statistics of the evaluation dataset are detailed in Table 1.

LVIS [16] dataset comprises 1,203 object categories, which significantly exceeds the number of categories in COCO. Following ViLD[14], we select 866 frequent and common categories as the training categories and reserve the remaining 337 rare categories as novel categories.

Evaluation Metrics. Following [24], we present the mean Average Precision (AP) for all categories in aggregate. Additionally, we report on AP_n for the novel category.

5.2. Implementation Details

In our implementation, the CLIP image encoder and text encoder are kept frozen, whereas the pixel decoder, transformer decoder, and mask prediction head undergo training. This strategy effectively harnesses the pre-trained capabilities of CLIP for mask classification and query matching, utilizing query embeddings, thereby obviating the need for category and identity annotations. The CLIP model employed in our framework adopts the ViTB/32 architecture for both the text and image encoders. Throughout the training process, the parameters of these encoders remain frozen. For optimization purposes, we utilize the AdamW optimizer, with the base learning rate set to 110^{-4}. $\tau = 0.05$. α_T, α_M, α_V, and α_S are all set to 1 based on grid parameter tuning experience. Following a step-wise learning rate schedule, this rate is subsequently reduced by a factor of 10 at 90% and 95% of the total training steps. To assess the robustness and generality of our proposed method, we conduct experiments using two distinct backbones: ResNet-50 and SwinB. Our experimental setup is analogous to that of OV2Seg [24]. All experiments are conducted on a high-performance machine equipped with eight NVIDIA A800 GPUs, each possessing 80GB of memory. To facilitate a fair comparison, we also present the inference speed of our method when executed on a single NVIDIA A100 GPU.

5.3. Main Results

The experimental results presented in Table 2 provide a comprehensive comparison of our proposed SWBridge method with several state-of-the-art approaches across multiple video instance segmentation datasets. A significant finding reveals that SWBridge exhibited superior performance compared to existing methods across 19 out of 22 evaluation metrics in various settings. Furthermore, it attained suboptimal results in the remaining three metrics, thereby underscoring its efficacy in addressing the inherent complexity associated with Open-Vocabulary VIS. This result not only verifies the effectiveness of SWBridge in its design, but also shows that it can adapt well to the characteristics of different datasets, such as changes in video length, number of instances, scene complexity, etc.

LV-VIS. When trained on the diverse and extensive LVIS dataset, which includes a large number of frequent and common categories, SWBridge exhibits robust generalization capabilities. This is particularly evident in its performance on the LV-VIS dataset, where it achieves the highest AP for both the validation and test sets, including the novel categories (AP_n). The superior performance on LV-VIS underscores SWBridge's ability to effectively recognize and segment objects from both base and novel categories, even in the presence of a significant number of unseen classes during training. Although powerful backbone networks such as ConvNeXt-B can enhance performance, we finds that the Swin-B backbone network, when combined with the SWBridge, yields even better results. This indicates that SWBridge has a synergistic effect with the backbone network, enabling more efficient capture of spatio-temporal features in videos, and the performance improvement exceeds that achieved by mere network upgrades.

YT-VIS. SWBridge demonstrates superior performance on the YT-VIS2019 and YT-VIS2021 datasets, further attesting to its strong capability in handling video instance segmentation tasks with a moderate number of categories. Remarkably, this method achieves competitive performance levels without the need for fine-tuning on specific video

Table 2. **Comparison with state-of-the-art methods.** All the methods are trained on image dataset LVIS and evaluated on the video instance segmentation datasets directly, i.e., they are not fine-tuned using the training set of each dataset.

Method	Backbone	LV-VIS val.		LV-VIS test		YT-VIS2019		YT-VIS2021		OVIS	BURST	
		AP	AP_n	AP	AP_n	AP	AP_n	AP	AP_n	AP	AP	AP_n
DetPro[10]-SORT[2]	R50	6.4	3.5	5.8	2.1	-	-	-	-	-	-	-
Detic[29]-SORT[2]	R50	6.5	3.4	5.7	2.1	14.6	3.5	12.7	3.1	6.7	1.9	2.5
DetPro[10]-OWTB[20]	R50	7.9	4.2	7.0	2.9	-	-	-	-	-	-	-
Detic[29]-OWTB[20]	R50	7.7	4.2	7.0	2.8	17.9	4.5	16.7	5.8	9.0	2.7	1.8
Detic[29]-XMem[7]	R50	8.8	5.4	7.7	3.6	-	-	-	-	-	-	-
OV2Seg[24]	R50	14.2	11.9	11.4	8.9	27.2	11.1	23.6	7.3	11.2	3.7	2.4
OVFormer[11]	R50+ViT-B	-	-	-	-	**34.8**	16.5	29.8	15.7	15.1	-	-
CLIP-VIS	R50	19.5	24.2	14.6	15.9	32.2	23.8	30.1	17.9	14.1	5.2	7.7
SWBridge(Our)	R50	**20.7**	**24.9**	**15.4**	**16.2**	33.4	**24.0**	**30.9**	**18.4**	**15.2**	**5.5**	**8.2**
Detic[29]-SORT[2]	SwinB	12.8	6.6	9.4	4.7	23.8	7.9	21.6	9.8	11.7	2.5	1.0
Detic[29]-OWTB[20]	SwinB	14.5	8.5	11.8	6.1	30.0	9.7	27.1	11.4	13.6	3.9	2.4
Detic[29]-XMem[7]	SwinB	16.3	10.6	13.1	7.7	-	-	-	-	-	-	-
OV2Seg[24]	SwinB	21.1	16.3	16.4	11.5	37.6	21.3	33.9	18.2	17.5	4.9	3.0
OVFormer[11]	SwinB+ViT-B	-	-	-	-	**44.3**	21.5	37.6	18.3	**21.3**	-	-
CLIP-VIS	ConvNeXt-B	32.2	40.2	25.3	30.6	42.1	27.5	37.9	22.0	18.5	**8.3**	12.7
SWBridge(Our)	SwinB	**33.0**	**40.5**	**26.0**	**31.0**	43.0	**28.0**	**38.5**	**22.5**	19.7	**8.3**	**12.8**

datasets. This characteristic indicates that SWBridge can effectively leverage image-based training data and successfully generalize to video data, thereby significantly reducing the reliance on extensive video-specific annotations. In comparison with state-of-the-art methods, although OVFormer achieves excellent results by incorporating an additional CLIP image encoder, SWBridge still outperforms CLIP-VIS in terms of the AP metric under similar model architectures. More notably, in the AP_n metric, which pertains to the segmentation of new class instances, our bridge strategy comprehensively surpasses the performance gains brought about by the CLIP encoder.

OVIS. Since the categories of the OVIS dataset highly overlap with the LVIS dataset used for training, we only reports the AP (average precision) performance metric. SWBridge even outperformed OVFormer with the extra CLIP on the R50 backbone and was second only to OVFoemer on the SwinB backbone. SWBridge's ability to maintain high accuracy under these conditions demonstrates its robustness in complex visual environments, where the visibility of objects changes significantly over time.

BURST. The BURST dataset, with its diverse and uncommon categories, poses a unique challenge due to the wide range of object types. SWBridge's competitive performance on this dataset, especially in terms of AP_n, suggests that it can effectively manage the intricacies of uncommon object categories, enhancing its applicability to real-world scenarios where encountering unseen objects is common.

Table 3. **Comparison of ablation results of bridge strategy.**

Variant	LV-VIS		OVIS	BURST	
	AP	AP_n	AP	AP	AP_n
Baseline	8.5	10.2	5.3	2.6	2.3
+ Spatio-Temporal Bridge	15.0	18.0	10.2	3.1	5.4
+ Text-Visual Bridge	19.9	23.3	14.2	5.3	7.9
+ $\mathcal{L}_{VT}$ in Eq. 10	**20.7**	**24.9**	**15.2**	**5.5**	**8.2**

5.4. Ablation Study

Our ablation experiments continue to use the LVIS dataset for training with ResNet50 as the Backbone and are evaluated on the validation sets of the LV-VIS, OVIS, and BURST datasets, aiming to measure the performance of the proposed SWBridge more comprehensively and rigorously.

Effectiveness of Bridge Strategy. We first established a baseline variant that removed the Spatio-Temporal Bridge bridge and Text-Visual Bridge modules, while discarding the additional contrast loss $\mathcal{L}_{VT}$ and retaining only the basic $\mathcal{L}_{Seg}$ loss function. In order to achieve instance association, we adopted the memory module proposed in OV2Seg to replace the original spatiotemporal bridge function.

The experimental results presented in Table 4 confirm the significant effectiveness of each component in the SWBridge method. Specifically, upon integrating the spatio-temporal bridge module into the baseline variant, we observed notable improvements in both AP and AP_n met-

Table 4. **Impact of ablation results of Importance Weighting.**

Variant	LV-VIS		OVIS	BURST	
	AP	AP_n	AP	AP	AP_n
w/o Weight	15.3	17.0	12.3	4.5	5.3
w/o Weight in §4.1	18.5	22.3	13.7	4.9	6.8
w/o Weight in §4.3	19.6	23.5	14.5	5.2	7.5
SWBridge(Our)	**20.7**	**24.9**	**15.2**	**5.5**	**8.2**

Table 5. **Impact of ablation results of Importance Weighting.**

Variant	LV-VIS		OVIS	BURST	
	AP	AP_n	AP	AP	AP_n
Huberized	18.8	22.4	13.8	4.7	6.4
Cosine	18.6	22.2	13.9	4.8	6.5
Wasserstein	19.5	23.3	14.4	5.1	7.0
Sliced Wasserstein	**20.7**	**24.9**	**15.2**	**5.5**	**8.2**

rics across all datasets. This outcome indicates that the spatio-temporal bridge module efficiently captures spatio-temporal correlations between video frames, thereby aiding the model in more accurately understanding and tracking the dynamic changes of instances within videos. Furthermore, the model's performance was significantly enhanced once again after adding the text-visual bridge module on top of the spatio-temporal bridge module. By introducing textual information, the text-visual bridge provides the model with rich semantic cues, greatly enhancing its accuracy in recognizing and understanding instances. Additionally, $\mathcal{L}_{VT}$ synergizes with other bridge strategy modules, collectively driving a comprehensive improvement in the model's performance and generalization.

Impact of SW Weighting Mechanism. Table 4 shows that without a weighting mechanism in the spatio-temporal or text-visual bridges, the model struggles to utilize key information from instances with marked feature differences, limiting performance. However, incorporating the weighting mechanism into both bridges notably improves model performance. This indicates the mechanism helps the model focus on crucial instances with significant feature differences, assigning them greater weights during distance calculation. This adjustment enhances the model's accuracy and efficiency in processing complex data.

Comparison of Distance Measurement. In Tab. 5, we compare distance metrics for SWBridge. Huberized distance is sensitive to large differences, Cosine similarity ignores feature magnitudes, and Wasserstein distance misses local details. Our SW distance comprehensively considers high-dimensional feature differences with importance weighting, making it robust for complex domain shifts.

Sensitivity analysis of L. In the bridge strategy, we ablated the number of sampling iterations to explore its relationship with model performance (Fig.2). Results show a non-linear relationship: increasing iterations significantly improves AP and AP_n across all datasets. However, excessive iterations introduce noise and computational burden, potentially causing overfitting or reducing training efficiency.

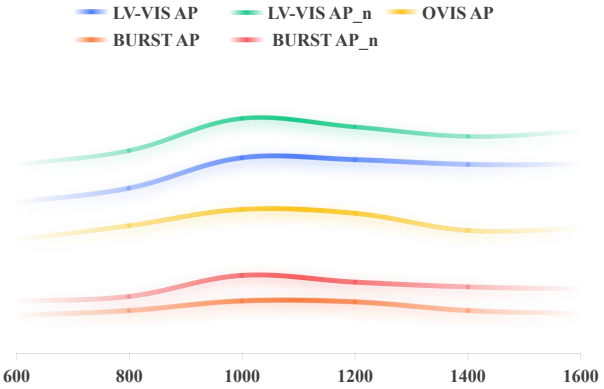

Figure 2. Ablation experiment on the number of random projection sampling L.

Table 6. **Comparison of accuracy, FLOPs, and inference speed on the validation set of the LV-VIS dataset.**

Method	Backbone	AP	FLOPs	FPS
OV2Seg[24]	R50	14.2	**238.2G**	26.1
CLIP-VIS [31]	R50	19.5	244.1G	31.4
SWBridge (Our)	R50	20.7	241.8	24.5
OV2Seg[24]	SwinB	21.1	448.2G	17.2
CLIP-VIS [31]	ConvNeXt-B	**32.2**	409.3G	21.0
SWBridge (Our)	SwinB	20.7	457.2	16.3

Complexity Interpretation. On the LV-VIS dataset (Tab. 6), SWBridge demonstrates unique advantages. Tested on an A100 GPU, it exhibits comparable or better accuracy than leading methods with R50 or SwinB backbones. Notably, the SwinB version has significantly higher FPS, attributed to our bridge strategy that achieves cross-domain consistency and inherits SW optimization efficiency.

6. Conclusion

In this paper, we propose an innovative Sliced Wasserstein Bridging Learning Framework (SWbridge), aimed at tackling the challenging task of Open-Vocabulary Video Instance Segmentation. Inspired by metric learning, our framework utilizes the Sliced Wasserstein distance as a core tool to effectively bridge the spatial-temporal and text-visual domains involved in the task. Through domain alignment, we successfully map features from different domains into a unified metric space, preserving temporal consistency and learning intrinsic consistent features across modalities. This significantly improves the fusion of text and visual information, addressing one of the core challenges in Open-Vocabulary VIS. Looking ahead, we plan to explore better optimization methods for the Sliced Wasserstein distance to further enhance the performance of our framework.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 2, 6

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 7

[3] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015. 2

[4] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *ECCV*, 2020. 1

[5] 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. 3

[6] 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. 2

[7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 7

[8] Zesen Cheng, Kehan Li, Hao Li, Peng Jin, Chang Liu, Xiawu Zheng, Rongrong Ji, and Jie Chen. Instance brownian bridge as texts for open-vocabulary video instance segmentation. *arXiv preprint arXiv:2401.09732*, 2024. 1, 2, 5

[9] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *ECCV*, 2024. 2

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 7

[11] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *ECCV*, 2024. 1, 2, 4, 7

[12] Damien Garreau, Rémi Lajugie, Sylvain Arlot, and Francis Bach. Metric learning for temporal sequence alignment. *NeurIPS*, 2014. 1

[13] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 2

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 6

[15] Pinxue Guo, Tony Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, Zhaoyu Chen, and Wenqiang Zhang. Openvis: Open-vocabulary video instance segmentation. *arXiv preprint arXiv:2305.16835*, 2023. 1, 2

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

[17] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and select: Semantic alignment metric learning for few-shot learning. In *ICCV*, 2019. 1

[18] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, 2022. 2

[19] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *CVPR*, 2023. 3

[20] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *CVPR*, 2022. 7

[21] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2, 6

[22] 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. 1, 2, 3

[23] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *ICCV*, 2023. 2

[24] 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, 4, 5, 6, 7, 8

[25] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 1

[26] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 6

[27] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *ICCV*, 2023. 1

[28] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *CVPR*, 2021. 2

[29] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 7

[30] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[31] 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. 1, 3, 4, 8