Semantic and Sequential Alignment for Referring Video Object Segmentation

Feiyu Pan[1] Hao Fang[1] Fangkai Li[1]
Yanyu Xu[1] Yawei Li[2] Luca Benini[2,3] Xiankai Lu[1]*
[1]Shandong University [2]ETH Zürich [3]University of Bologna
https://github.com/tavarich/SSA

Abstract

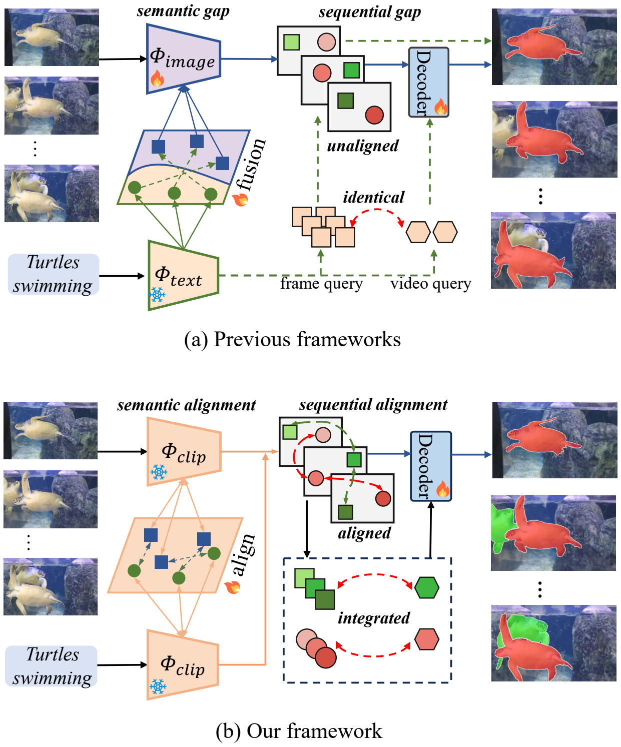
*Referring video object segmentation (RVOS) seeks to segment the objects within a video referred by linguistic expressions. Existing RVOS solutions follow a "fuse then select" paradigm: establishing semantic correlation between visual and linguistic feature, and performing frame-level query interaction to select the instance mask per frame with instance segmentation module. This paradigm overlooks the challenge of semantic gap between the linguistic descriptor and the video object as well as the underlying clutters in the video. This paper proposes a novel **S**emantic and **S**equential **A**lignment (SSA) paradigm to handle these challenges. We first insert a lightweight adapter after the vision language model (VLM) to perform the semantic alignment. Then, prior to selecting mask per frame, we exploit the trajectory-to-instance enhancement for each frame via sequential alignment. This paradigm leverages the visual-language alignment inherent in VLM during adaptation and tries to capture global information by ensembling trajectories. This helps understand videos and the corresponding descriptors by mitigating the discrepancy with intricate activity semantics, particularly when facing occlusion or similar interference. SSA demonstrates competitive performance while maintaining fewer learnable parameters.*

1. Introduction

Referring video object segmentation (RVOS) is an emerging multimodal video understanding task, which aims at segmenting the target from a video based on linguistic description [7, 10, 18, 47]. It has diverse applications in fields such as human-robot interaction [40, 41] and video editing [3], garnering attention from the academic community.

Traditional RVOS approaches [4, 54] typically focus on simple yet static scenes with prominent object, where one static expression referrs to only one target [10, 18, 47]. Such isolated objects can be easily captured by static attributes in a single frame. However, in real-world video scenarios, existing pipelines struggle to seamlessly interpret the motion and multi-targets expressions with complex and dynamic scenarios due to substantial movement, abrupt scene change and non-salient targets [7, 26]. This leads to the challenge of aligning both video-text and trajectory-instance relations.

RVOS in real-world environment requires explicitly identifying the particular target trajectory referred by the motion guidance. This underscores the necessity for robust visual-text semantic understanding to establish the corre-



(a) Previous frameworks

(b) Our framework

Figure 1. (a) **Previous frameworks** extracted features from independent extractors, while not effectively leverage the consistency of the object query sequence. (b) **Our SSA framework** yields sufficient visual-linguistic interaction through semantic alignment by Vision-Language Model and global video modeling via sequential alignment.

[*] Corresponding author: *Xiankai Lu*.

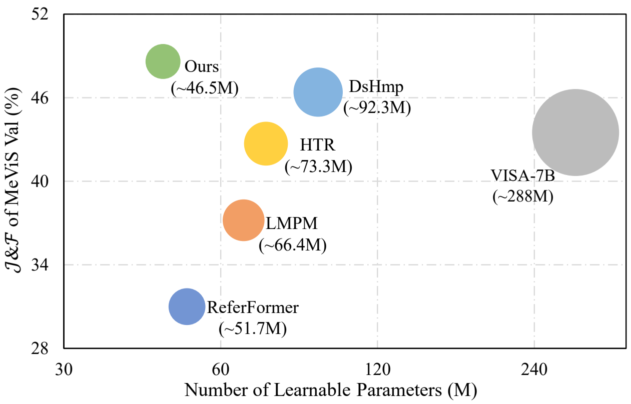


Figure 2. Comparison of correlation between performance and trainable parameters on MeViS dataset [7]. Circle size represents the numbers of trainable parameters for each model.

spondence between video content and motion description, as well as the capacity to comprehend the entire video that ensure consistent object sequence segmentation. Existing methods [4, 7, 12, 13, 53, 54] utilize independent backbones to learn visual-text relationships from scratch, which may cause semantic gap during training (Fig. 1 (a)). After the challenging relationship modeling, they segment the target for each frame individually [4, 53, 54] or introduce video query [14] initialized solely by text feature, then interacting with heavily processed frame contents to generate sequence output [7, 13], both of which are suboptimal due to overlooking the comprehensive video representation from object sequences. To enhance the solution in real-world scenarios, a paradigm shift is needed, where both multi-modality semantic features and sequential information are uniformly aligned to improve multi-modal understanding and consistent temporal perception.

To alleviate the misalignment issues in real-world scenes overlooked by pervious methods, we propose a novel framework called SSA (Fig. 1 (b)), aimed at enhancing the **semantic alignment** across visual-text features and **sequential alignment** across frame contents. Specifically, we leverage a frozen Vision-Language Model (VLM) to provide semantically correlated visual-textual features. We regard RVOS as a complex concept that involves various fine-grained entities, such as motions, scenes, and instances parsing. So we further propose a lightweight adapter to refine the visual-text semantic alignment between pixel-level video content and referring description for sufficient interactions, as we hypothesize that, the textual representations of the relevant entities should be aligned with specific regions of the image.

Furthermore, to improve the global understanding of the entire video context, we incorporate a sequential alignment module to enhance the trajectory-to-instance correspondence. By ensembling these global trajectories, we can obtain a video-level context about the whole video thereby facilitating the consistent mask generation. This synergistic integration of sequential alignment and semantic alignment further improves the overall performance.

The contributions of this paper are as follows:

- We propose the **S**emantic and **S**equential **A**lignment (SSA) framework that leverages video-to-text correspondences as well as trajectory-to-instance correspondence to guide the adaptation from VLM to RVOS.
- We propose a new lightweight adapter to refine the visual-language alignment of VLM. It encourages SSA to extract the semantics of the most important image regions given the referring descriptor and alleviate training difficulties.
- We further conduct sequential alignment on the trajectory to fully exploit global context, improving the overall interpretation of the video.
- SSA achieves new state-of-the-art results, especially on the challenging MeViS dataset [7] by **2.2%**$\mathcal{J}\&\mathcal{F}$ with fewer trainable parameters, and exhibits significant improvement compared to LVLM methods (Fig. 2).

2. Related Work

Referring Video Object Segmentation. Referring Video Object Segmentation (RVOS) requires segmenting the target within a video based on the corresponding language expression [10, 24]. Seo et al. [47] establish the first large-scale RVOS benchmark, driving progress in the filed of RVOS. Many prior approaches in referring video segmentation have primarily adapted VIS [9, 56] (*e.g.* VisTR [51]) as streamlined pipeline, such as MTTR [51] and ReferFormer [54]. To better modeling temporal context [32], subsequent explorations include OnlineRefer [53] that leverages query propagation to explore online RVOS, TempCD [49] that interacts with a global refer token with sequences of intra-frame queries.

Recently, a new large-scale benchmark called MeViS [7] is introduced. Compared to the previous RVOS datasets [10, 18, 47], MeViS emphasizes the motion attributes of videos with a large variety of motion expressions, as well as flexible expressions that refer to multi-targets. Thus, most of the current methods [4, 12, 47, 49, 53, 54, 60] can not handle this complex setting well, which reflects real-world scenarios. To address this new issue, a “fuse then select” paradigm has been raised. Typical works include LMPM [7] and DsHmp [13]. These methods first employ an advanced instance segmentation model to segment the possible objects according to fusing the text features. Then, a decoder module works for selecting the referred object and performing heavily cross-frame association. Despite handling multiple instances segmentation successfully, these methods either yield suboptimal performance [7] or have complex network structures [13].

Although our approach also builds upon query-based segmentation architecture, it differs in key aspects. We

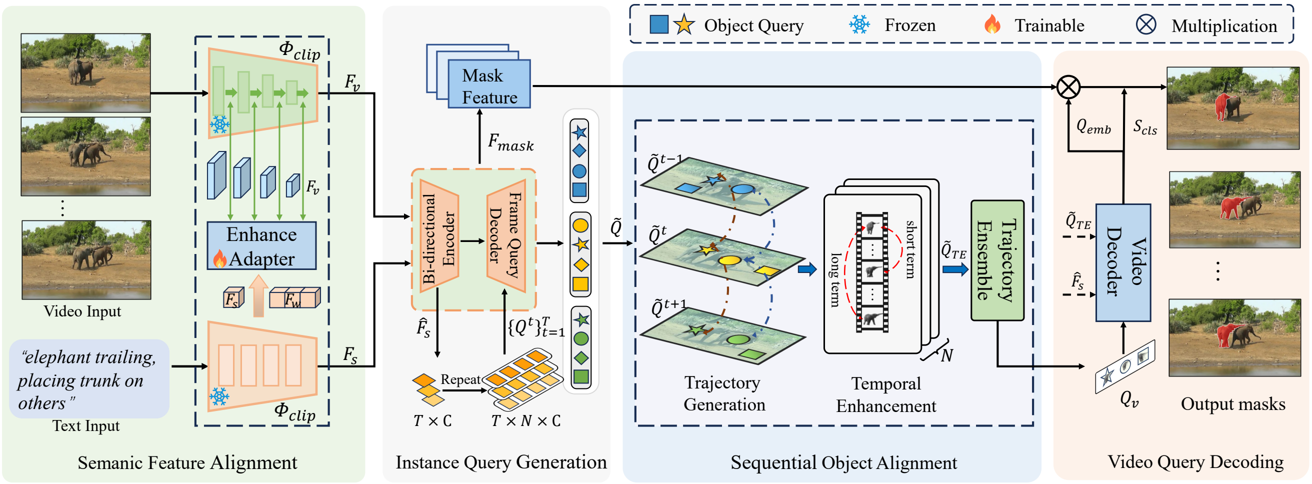


Figure 3. The overview architecture of the proposed method. It primarily contains Semantic Feature Alignment Encoders (Sec. 3.2) to capture semantic aligned visual-linguistic features, Instance Query Generation Module for extracting object-centric representation, and Sequential Object Alignment Module (Sec. 3.3) to align instance queries across the timeline for global video understanding.

revisit RVOS from a multi-modal perception perspective and decompose it into video-to-text correspondences and trajectory-to-instance correspondence. Correspondingly, we propose semantic alignment and sequential alignment modules. First, this perspective enhances the model's ability to associate textual queries with relevant video content. Second, the alignment paradigm enables precise and consistent sequence-instance modeling, leading to better results.

Vision-Language Models. Vision-Language Models (VLM) learn from image and text simultaneously, demonstrating strong cross-modal reasoning and generalization capabilities [2, 21, 23, 42, 46, 61]. VLMs typically use contrastive learning [16, 37, 42], generative methods [45, 48, 57] and alignment objectives [22, 27] to achieve cross-modal pre-training. With the powerful pre-trained knowledge of VLMs, they are now widely applied to various visual analysis tasks [19, 52, 59]. In the video domain, early works [9, 11, 50] utilize CLIP [42] as knowledge prior to achieve VIS in open vocabulary scenarios. Meanwhile, recent works, such as VISA [55] and VideoLISA [1], explore large VLMs (LVLMs) to enable textual reasoning and referring capabilities for video segmentation. Nevertheless, typical RVOS models use standalone vision-text encoders, leading to training difficulties caused by feature misalignment. Additionally, the current LVLM-based solutions [1, 55] and SAM2-based attempts [15, 43, 44] can be costly and yields suboptimal results on the challenging RVOS benchmark. In contrast, inspired by the success of VLM's fine-tuning trend, we utilize frozen VLM (CLIP [42]) as powerful feature extractors and introduce a context-aware adapter to alleviate the training challenges of RVOS.

3. Methods

3.1. Revisiting "Fuse then Select" Paradigm

Given a video clip with T frames $\mathcal{I} = \{I_t\}_{t=1}^T$ and a referring expression $\mathcal{E} = \{e_n\}_{n=1}^N$ consisting of N words, RVOS needs to generate the binary masks $\mathcal{M} = \{m_t\}_{t=1}^T$, $m_t \in \mathbb{R}^{H\times W}$ of the targets guided by the expression $\mathcal{E}$. Before elaborating on our method (§3.2), we first give a brief introduction to "fuse then select" paradigm in RVOS.

Existing approaches [4, 12, 13, 49, 53, 54] often utilize pretrained vision encoder and language encoder to extract visual features $\mathcal{F}_v$ and textual feature $\mathcal{F}_s$ independently:

$$\begin{aligned}\mathcal{F}_v &= \Phi_{image}(\mathcal{I}) \in \mathbb{R}^{T\times C_i\times H_i\times W_i},\\ \mathcal{F}_s &= \Phi_{text}(\mathcal{E}) \in \mathbb{R}^{1\times C_l},\end{aligned} \tag{1}$$

where Φ_{image} and Φ_{text} denote Swin Transformer [29] and RoBERTa [28], separately. $i \in [1, 4]$ is the stage index of the image encoder, C_l is the dimension of textual features.

After that, "fuse then select" methodology [7, 13] employs a decoder to fuse the visual and textual features together, and generate language-guided object embeddings which correspond to the potential objects. Then, these potential objects are associated into the trajectory candidates. Finally, the textual features are reused to select the predicted object trajectories from the candidates.

3.2. Semantic Feature Alignment

In contrast to the prior RVOS methods, we propose a new framework named SSA that learns **S**emantic and **S**equential **A**lignment for RVOS. Specifically, it fully exploits both the semantic correspondence between text and visual features, and the temporal alignment within video sequences. Building on this, our approach enables us to construct an elegant

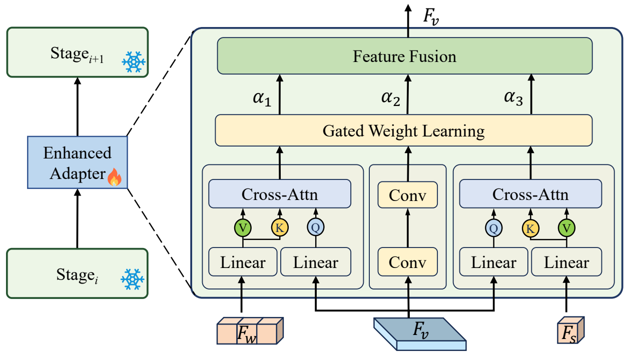


Figure 4. Illustration of the Semantic Alignment Refine Module.

model that is both strong in performance and effective in training. The overview of our approach SSA is depicted in Fig. 3.

To obtain more accurate grounding region according to the text embeddings, we first devise a semantic feature alignment in a progressive manner.

Coarse Feature Alignment. To address the rich semantic diversity in real-world RVOS scenarios, we firstly leverage frozen Vision-Language Model (*i.e.*, CLIP [42]) to obtain semantic correlated visual-linguistic representations jointly:

$$(\mathcal{F}_v, \mathcal{F}_s, \mathcal{F}_w) = \Phi_{clip}(\mathcal{I}, \mathcal{E}), \tag{2}$$

where Φ_{clip} denotes the CLIP model and $\mathcal{F}_w \in \mathbb{R}^{N \times C_l}$ denotes textual features of each word.

Compared to Eq. (1), Eq. (2) facilitates the model's comprehensive understanding of both the video and expression while reducing training cost.

Semantic Alignment Refine Module. Considering CLIP is pre-tained with large-scale image-text pairs, it may not handle fine-grained visual perception tasks well, such as image segmentation [17]. Therefore, we further design a lightweight adapter (as shown in Fig. 4) for each stage of CLIP to enhance vision-language semantic alignment, promoting a refined grounding of relevant entities according to the referring sentences.

Due to referring expression provide fine-grained representations of the sentence, and we argue that they can ground the corresponding regions in video frames. Thus the proposed adapter consists of two-granularity cross-attention among visual features and textual features: {visual feature $\mathcal{F}_v$ and *sentence-level* textual features $\mathcal{F}_s$}, {visual feature $\mathcal{F}_v$ and *word-level* textual features $\mathcal{F}_w$}. In addition, to preserve the general knowledge of CLIP, we also utilize convolution blocks to build residual path:

$$\begin{aligned} f_{mw} &= \text{softmax}(\frac{\mathcal{F}_v \cdot \mathcal{F}_w}{\sqrt{C}})\mathcal{F}_w, \\ f_{ms} &= \text{softmax}(\frac{\mathcal{F}_v \cdot \mathcal{F}_s}{\sqrt{C}})\mathcal{F}_s, \\ f_{mr} &= \Phi_{conv}(\mathcal{F}_v) + \mathcal{F}_v, \end{aligned} \tag{3}$$

where f_{mw}, f_{ms} and f_{mr} mean visual outputs from three streams, respectively. C denotes dimension of the joint embedding space, Φ_{conv} indicates convolution operation.

Through Eq. (3), the adapter generates refined visual-textual aligned features from multiple streams. To balance the semantics introduced by different feature streams, we employ an MLP serving as gating mechanism to learn dynamic weights for each stream. The final output at each stage is as follows:

$$\begin{aligned} \mathcal{W}_\alpha &= \text{softmax}(\text{MLP}([g(f_{mw}), g(f_{mr}), g(f_{ms})])), \\ \mathcal{F}_v &= \mathcal{W}_{\alpha 1} \cdot f_{mw} + \mathcal{W}_{\alpha 2} \cdot f_{mr} + \mathcal{W}_{\alpha 3} \cdot f_{ms}, \end{aligned} \tag{4}$$

where $g(\cdot)$ denotes global average pooling, $\mathcal{W}_\alpha \in \mathbb{R}^{3 \times 1}$ is the weight of different streams, $[\cdot]$ is the concatenation operation. Through the proposed adapter, multi-granularity textual features progressively align within the visual feature, while retaining VLM's powerful knowledge.

3.3. Sequential Object Alignment

Instance Query Generation. Once fused features $\mathcal{F}_v$ is obtained from VLM, the existing RVOS solutions often employ object query to serve as implicit representation of the target in the video. Typically, they are either initialized randomly for each frame [4] or sorely generated via sentence-level textual feature (Eq. (1)) [7, 12, 53, 54].

Considering the variations in the position and shape of objects between different video frames over time, we propose an object query generating mechanism to ensure that queries have variability across frames via cross-modal feature interaction (as depicted in Fig. 3).

In concrete, we firstly leverage 1×1 convolution block and linear layer to project the visual and textual features into the joint space, respectively. Subsequently, bidirectional cross-attention operation is adopted at bi-directional encoder to conduct interaction between visual and textual features frame by frame:

$$\begin{aligned} \hat{\mathcal{F}}_v^t &= \text{softmax}(\frac{\mathcal{F}_v^t W_{t2v}^Q \cdot (\mathcal{F}_s W_{t2v}^K)^\intercal}{\sqrt{C}})\mathcal{F}_s W_{t2v}^V, \\ \hat{\mathcal{F}}_s^t &= \text{softmax}(\frac{\mathcal{F}_s W_{v2t}^Q \cdot (\mathcal{F}_v^t W_{v2t}^K)^\intercal}{\sqrt{C}})\mathcal{F}_v^t W_{v2t}^V, \end{aligned} \tag{5}$$

where $W^Q_{\{t2v,v2t\}}$, $W^K_{\{t2v,v2t\}}$, $W^V_{\{t2v,v2t\}}$ are learnable projection weights, t denotes frame index. $\hat{\mathcal{F}}_s^t \in \mathbb{R}^{1 \times C}$ is updated textual feature of t-th frame. Additionally, $\mathcal{F}_v$ is utilized to generate mask feature $\mathcal{F}_{mask}$, which is used to produce the final mask sequence.

Instance queries $\{Q^t\}_{t=1}^T, Q^t \in \mathbb{R}^{N \times C}$ generated by repeating $\hat{\mathcal{F}}_s$ for N times, form as initialized sequences. Then, $\{Q^t\}_{t=1}^T$ are fed into the frame query decoder to learn the potential target representation $\{\tilde{Q}^t\}_{t=1}^T$ for each frame.

Instance Trajectory Generation. Object queries $\{\tilde{Q}^t\}_{t=1}^T$ are extracted independently for each frame, which may lead

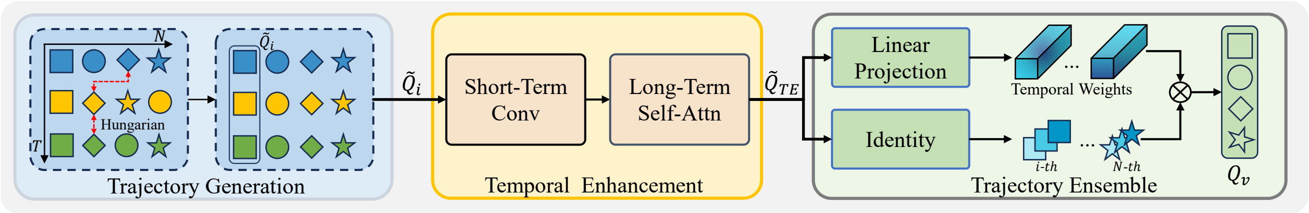


Figure 5. Illustration of Sequential Object Alignment Module.

to misalignment for the same instance along the temporal dimension. To address this issue, we utilize Hungarian matching algorithm [20] to match $\tilde{Q}$ of adjacent frames along timeframes to model trajectories for each potential instance (as shown in Fig. 5).

Upon establishing trajectories for $\{\tilde{Q}_i | i \in [1, N], \tilde{Q}_i \in \mathbb{R}^{T \times C}\}$, we leverage convolution block for short-term information and self-attention layer for global view for trajectory's temporal enhancement:

$$\begin{aligned} \text{ShortTerm}(\tilde{Q}_i) &= \Phi_{Conv1D}(\tilde{Q}_i) \in \mathbb{R}^{T \times C}, \\ \text{LongTerm}(\tilde{Q}_i) &= \Phi_{TempAttn}(\tilde{Q}_i) \in \mathbb{R}^{T \times C}, \end{aligned} \tag{6}$$

where $\tilde{Q}_i$ is the i-th instance of N trajectories, Φ_{Conv1D} and $\Phi_{TempAttn}$ denote 1D convolution and temporal attention.

Trajectory to Instance Correspondence. Once obtained the temporal enhanced instance trajectories $\tilde{Q}_{TE}$, we can ensemble it as a global description of the whole video. The global representation of instance (video query) is then obtained by temporal weighted fusion:

$$\begin{aligned} \mathcal{A} &= \text{softmax}(\Phi_{FC}(\tilde{Q}_{TE})) \in \mathbb{R}^{T \times N \times 1}, \\ Q_v &= \sum_{t=1}^{T} \mathcal{A}^t \cdot \tilde{Q}_{TE}^t \in \mathbb{R}^{N \times C}, \end{aligned} \tag{7}$$

where Φ_{FC} denotes linear projection layer with learnable weight in $\mathbb{R}^{1 \times C}$, $\mathcal{A}$ is the temporal weight in each frame through the softmax function.

Compared to previous methods that utilize frame queries directly for sequence output [4, 12, 53, 54] or rely on solely text features as global video query [7, 13, 33], the proposed approach integrates global modeling from the entire video that sufficiently mines the global information.

3.4. Video Query Decoding

Similar to previous works [7, 13, 14], given the video queries Q_v, frame queries $\tilde{Q}_{TE}$ and language features $\hat{\mathcal{F}}_s$, video decoder works for generating binary classification scores S_{cls} and mask embeddings Q_{emb}, separately.

Then Q_{emb} is filtered through S_{cls} by argmax operation for single object scenarios [38, 47] or threshold operation for multiple instance scenarios [7]. Then, it multiplies with mask feature $\mathcal{F}_{mask}$ to obtain the final mask sequence of the identified target(s).

3.5. Model Implementation and Training

Model Implementation. Specifically, we use ConvNeXt [30] based CLIP [42] as feature extractors (§ 3.2), as ConvNeXt naturally extracts multi-scale feature maps from image, making it suitable for segmentation tasks. We utilize modified mask2former [5] to implement instance query generation and transformer blocks for video query decoder in §3.3.

Training. Following [7, 13, 14], we employ the $\mathcal{L}_f$ to calculate the loss between per-frame predictions and frame-wise annotations, along with $\mathcal{L}_v$ as video-level supervision. We implement the $\mathcal{L}_f$ and $\mathcal{L}_v$ by binary cross-entropy loss with dice loss. The overall training objective is given by:

$$\mathcal{L}_{train} = \mathcal{L}_v + \mathcal{L}_f + \lambda_{sim}\mathcal{L}_{sim}, \tag{8}$$

where λ_{sim} is the weight for the similarity loss $\mathcal{L}_{sim}$ [14].

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. The proposed methods are evaluated on three popular RVOS datasets: Ref-Youtube-VOS [47], Ref-DAVIS17 [18] and MeViS [7]. Ref-Youtube-VOS [47] establishes the first large-scale dataset consisting of 3,978 videos with approximate 15k language descriptions and 202 video samples are used for online testing. Ref-DAVIS17 [18] contains 90 videos modified DAVIS17 [39] by providing additional textual annotations, which allocates 30 videos for validation. MeViS [7] is a new large-scale benchmark that highlights motion information analysis in video. It includes 2,006 video sequences with 443k fine-grained segmentation annotations, and 28,570 sentences describing 8,171 instances in complex video scenarios. The *Val* dataset of MeViS contains 140 video for online evaluation, while *Valu* includes 50 videos for local validation.

We further validate on two additional datasets [10], and the details are provided in the supplementary materials.

Evaluation Metrics. Following previous works [7, 54], jaccard index $\mathcal{J}$ (region similarity), mean boundary similarity $\mathcal{F}$ (contour accuracy), and their average $\mathcal{J}\&\mathcal{F}$ are employed as evaluation metrics.

4.2. Experimental Settings

MeViS. For MeViS dataset, following previous works [7, 13], without pre-training on the image-level referring

Table 1. **Comparison with state-of-the-art methods on MeViS [7] *Val* and *Valu* dataset.** † denotes our model is trained on MeViS [7] and Ref-Youtube-VOS [7] jointly. See §4.3 for details.

Methods	Reference	*Val*			*Valu*		
		$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
Compared with expert methods							
URVOS [47]	ECCV'2020	27.8	25.7	29.9	-	-	-
LBDT [8]	CVPR'2022	29.3	27.8	30.8	-	-	-
MTTR [4]	CVPR'2022	30.0	28.8	31.2	-	-	-
ReferFormer [54]	CVPR'2022	31.0	29.8	32.2	-	-	-
VLT+TC [6]	TPAMI'2022	35.5	33.6	37.3	-	-	-
HTR [36]	TCSVT'2024	42.7	39.9	45.5	-	-	-
LMPM [7]	ICCV'2023	37.2	34.2	40.2	40.2	36.5	43.9
DsHmp [13]	CVPR'2024	46.4	43.0	49.8	55.3	51.0	60.4
Ours	-	**48.6**	**44.0**	**53.2**	**56.9**	**51.7**	**62.2**
Compared with LVLM methods							
VideoLISA-3.8B [1]	NeurIPS'2024	44.4	41.3	47.6	-	-	-
VISA-7B [55]	ECCV'2024	43.5	40.7	46.3	51.2	48.0	54.4
Ours †	-	**48.9**	**44.3**	**53.4**	**57.8**	**52.3**	**63.1**

datasets RefCOCO/+/g [34, 58], we train the model on MeViS directly. The optimization is performed with AdamW [31] for 50k iterations, with a learning rate initialized to $5e^{-4}$.

Ref-Youtube-VOS and Ref-DAVIS17. For these two datasets, following [13], we first conduct image-level pretraining on the RefCOCO/+/g datasets [34, 58]. After that, we fine-tune the model with videos from the Ref-Youtube-VOS. During the pre-training phase, the model is trained for 100k iterations. In the video-level training phase, we trained the model for 50k iterations. For Ref-DAVIS17, we directly apply the learned model from Ref-YouTube-VOS to Ref-DAVIS17 [18] without online fine-tuning.

Joint Video Datasets Co-training. Considering current VLM methods [1, 55] tend to be trained on image and video datasets jointly, we adopt a similar strategy to explore the performance potential by training our model on MeViS and Ref-Youtube-VOS jointly during the video training phase.

Table 2. **Comparison with SAM2 models on MeViS [7] *Val* dataset.** ‡ Results are adopted from [15]. See §4.3 for details.

Methods	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
Grounded-SAM 2 [44]‡	38.9	35.7	42.1
AL-Ref-SAM 2 [15]‡	42.8	39.5	46.2
Ours	**48.6**	**44.0**	**53.2**

4.3. Main Results

MeViS. In Tab. 1, we validate the proposed approach on the newly released motion expressions guided video segmentation dataset MeViS [7]. With comparable training costs, SSA outperforms other leading expert methods, surpassing the main counterpart DsHmp [13] by **2.2%** $\mathcal{J}\&\mathcal{F}$ on *val* set and 1.6% on *valu* set (**2.5%** on *val* and *valu* under our co-training strategy).

Additionally, in comparison with methods leveraging large vision-language models, SSA exceeds the contemporary methods and establishes a new state-of-the-art performance in terms of mean $\mathcal{J}\&\mathcal{F}$ (**48.9%**) *val* set and $\mathcal{J}\&\mathcal{F}$ (57.8%) on *valu* set. Notably, our method shows a marked performance gain over the best method VideoLiSA [1] by **4.5%**$\mathcal{J}\&\mathcal{F}$. These results demonstrate the importance of semantic alignment and sequential alignment. Furthermore, as shown in Fig. 2, our approach improves performance while reducing approximate 46M trainable parameters compared to current state-of-the-art DsHmp [13], alleviating the training difficulties for RVOS task.

Moreover, the recently introduced Segment Anything Model 2 (SAM2) [43] has achieved breakthrough progress in VOS tasks, and some works have migrated it to the RVOS domain [15, 44]. For experimental completeness, we also conduct experimental comparisons with SAM2 based pipelines. Nevertheless, as shown in Tab. 2, our method behaves better than the existing SAM2 attempts in the complex MeViS scenarios with a distinct performance gap: (AL-Ref-SAM 2 [15]: 48.6 *vs.* 42.8, 44.0 *vs.* 39.5, 53.2 *vs.* 46.2) and (Grounded-SAM 2 [44]: 48.6 *vs.* 38.9, 44.0 *vs.* 35.7, 53.2 *vs.* 42.1).

Ref-YouTube-VOS. The results on the Ref-YouTube-VOS dataset are presented in Tab. 3. Our method exceeds existing approaches on Ref-Youtube-VOS across all metrics by achieving 64.3% $\mathcal{J}\&\mathcal{F}$, which is **0.7%** higher than DsHmp.

Ref-DAVIS17. On Ref-DAVIS17 (Tab. 3), our approach achieves 67.3% $\mathcal{J}\&\mathcal{F}$ and consistently outperforming the top method (*i.e.*, DsHmp) by more than **3.3%** in terms of $\mathcal{J}\&\mathcal{F}$. These experiment results exhibit the generalization

Table 3. **Comparison with state-of-the-art models on Ref-Youtube-VOS [47] and Ref-DAVIS17 [18] datasets.** See §4.3 for details.

Methods	Reference	Backbone	Ref-Youtube-VOS			Ref-DAVIS17		
			$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
LBDT [8]	CVPR'2022	ResNet-50	49.4	48.2	50.6	54.3	-	-
MTTR [4]	CVPR'2022	Video-Swin-T	55.3	54.0	56.6	-	-	-
ReferFormer [54]	CVPR'2022	Video-Swin-T	59.4	58.0	60.9	59.6	56.5	62.7
OnlineRefer [53]	ICCV'2023	Video-Swin-B	62.9	61.0	64.7	62.4	59.1	65.6
HTML [12]	ICCV'2023	Video-Swin-T	61.2	59.5	63.0	-	-	-
R2VOS [25]	ICCV'2023	Video-Swin-T	61.3	59.6	63.1	-	-	-
SgMg [35]	ICCV'2023	Video-Swin-T	62.0	60.4	63.5	61.9	59.0	64.8
TempCD [49]	ICCV'2023	Video-Swin-T	62.3	60.5	64.0	62.2	59.3	65.0
SOC [33]	NeurIPS'2023	Video-Swin-T	62.4	61.1	63.7	63.5	60.2	66.7
LoSh [60]	CVPR'2024	Video-Swin-T	63.7	62.0	65.4	62.9	60.1	65.7
DsHmp [13]	CVPR'2024	Video-Swin-T	63.6	61.8	65.4	64.0	60.8	67.2
Ours	-	CLIP	**64.3**	**62.2**	**66.4**	**67.3**	**64.0**	**70.7**

capability of our method.

4.4. Ablation Studies

Table 4. **Ablation study on main components of SSA.** *Semantic* indicates Semantic Feature Alignment, *Sequential* denotes Sequential Object Alignment. See §4.4 for details.

Components		$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
Semantic	Sequential			
✗	✗	42.4	37.4	47.4
✓	✗	46.3	41.5	51.1
✗	✓	45.1	40.8	49.4
✓	✓	**48.6**$_{\uparrow 6.2}$	**44.0**$_{\uparrow 6.6}$	**53.2**$_{\uparrow 5.8}$

In this section, we conduct ablation studies on the challenging dataset MeViS [7] to investigate our proposed approach. By removing the proposed core components, we build vanilla baseline that shares a similar structure with LMPM [7].

Table 5. **Comparison with DsHmp [13] variant models on MeViS dataset.** See §4.4 for details.

Methods	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
DsHmp [13]	46.4	43.0	49.8
+ CLIP Encoders (Eq. (2))	47.2	42.9	51.5
+ Correspond (Eq. (7))	47.6	43.4	51.8
Ours	**48.6**	**44.0**	**53.2**

Component Analysis. We first verify the effectiveness of main components of our framework. As shown in Tab. 4, semantic feature alignment mechanism provides an significant performance gain by 3.9 % of $\mathcal{J}\&\mathcal{F}$ compared to baseline who utilizes Swin [29] and RoBERTa [28] as encoders (2^{nd} row). This indicates the necessity of semantic alignment in RVOS modeling.

Adding sequential object alignment to baseline improves the performance by 2.7% of $\mathcal{J}\&\mathcal{F}$ (3^{rd} row). Finally, after integrating all the main components together, we observe a substantial improvement up to 6.2% of $\mathcal{J}\&\mathcal{F}$ compared to the baseline (last row).

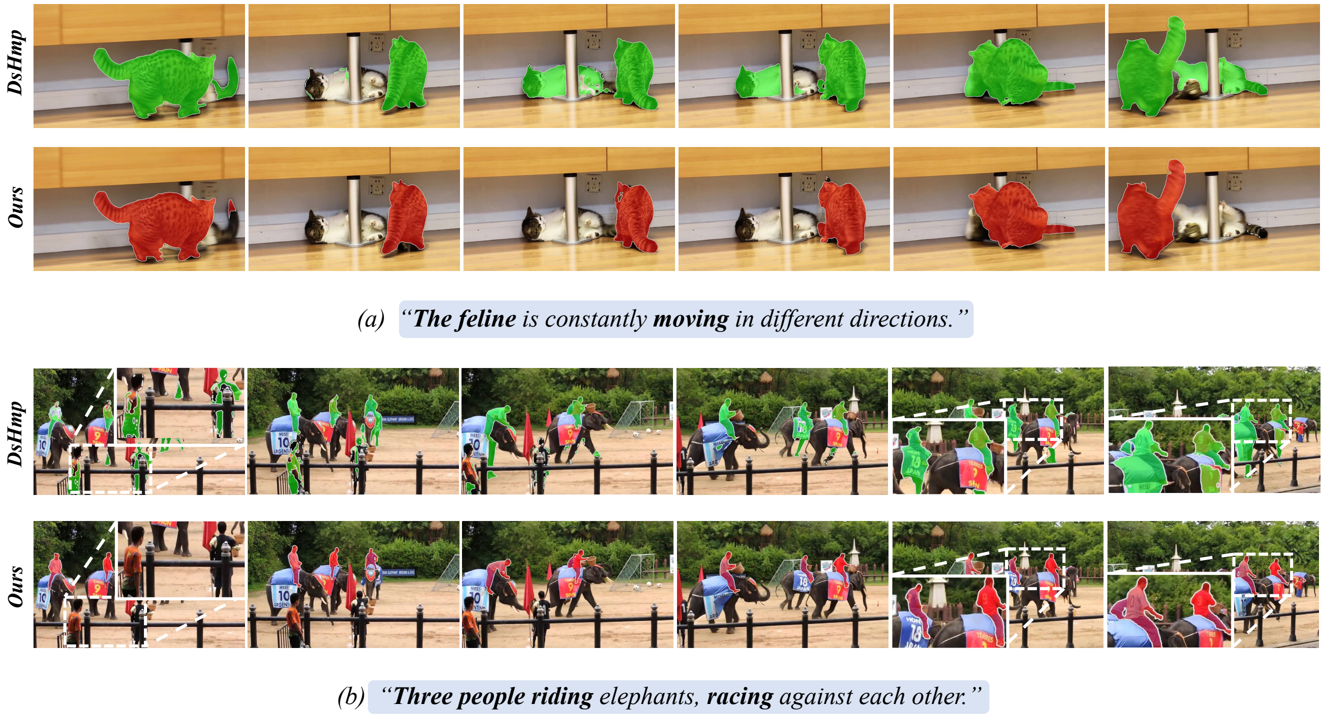
Additionally, we extend the DsHmp [13] to validate the fairness. As shown in Tab. 5, when employing the same CLIP encoders, DsHmp slightly improves by 0.8%, but still remains 1.4% lower than SSA. Moreover, adding instance correspondence (Eq. 7) further improves the valina Dshmp by 1.2%. These analysis demonstrate the effectiveness as well as suitability of the proposed components.

Table 6. **Comparisons of different Semantic Feature Alignment strategies.** *Frozen-CLIP* denotes freezing the entire CLIP model, *Fine-tuning* indicates fine-tune the last stage of CLIP, *Adapter* refers to the proposed adapter. See §4.4 for details.

Frozen-CLIP	Fine-tuning	Adapter	Trainable-params ↓	$\mathcal{J}\&\mathcal{F}\uparrow$
✗	✗	✗	72.3M	45.1
✓	✗	✗	**39.4M**	47.5
✓	✓	✗	100.6M	<u>47.9</u>
✓	✗	✓	<u>46.5M</u>	**48.6**

Effectiveness of Adapter. Further ablation studies on Semantic Feature Alignment is shown in Tab. 6. Compared to the baseline, it can be seen 2.4% $\mathcal{J}\&\mathcal{F}$ performance improvement is achieved by introducing frozen CLIP model, while reduce approximately 46% trainable parameters. Compared with fine-tuning the last stage of ConvNext, the training cost is increased but bring a slight performance gain of 0.4%. Finally, when adopting the proposed adapter, SSA further improves the $\mathcal{J}\&\mathcal{F}$ by 1.1%, with affordable extra parameters increase than fine-tuning.

In-depth Analysis of Sequential Object Alignment. Finally, we study the components of Sequential Object Alignment. As shown in Tab. 7, by removing bi-directional attention in Eq. (5) (2^{nd} row) for instance query generation,



(a) *"The feline is constantly moving in different directions."*

(b) *"Three people riding elephants, racing against each other."*

Figure 6. Qualitative Results on Mevis dataset [7]. The first row illustrates the segmentation results of DsHmp [13] while the second row indicates the segmentation results of ours.

Table 7. **Ablation study on Sequential Object Alignment.** Sequential indicates the overall module, the rest rows indicate the variants by removing specific components. See §4.4 for details.

Architecture	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
Sequential	**48.6**	**44.0**	**53.2**
w/o Bi-directional	$47.8_{\downarrow 0.8}$	$43.2_{\downarrow 0.8}$	$52.3_{\downarrow 0.9}$
w/o Ensemble	$47.2_{\downarrow 1.4}$	$42.6_{\downarrow 1.4}$	$51.9_{\downarrow 1.3}$
w/o Trajectory/Ensemble	$46.1_{\downarrow 2.5}$	$41.2_{\downarrow 2.8}$	$51.0_{\downarrow 2.2}$

the segmentation performance drops 0.8% of $\mathcal{J}\&\mathcal{F}$. Without trajectory ensemble (3^{rd} row) in Eq. (7) for instance correspondence, we observe a 1.4% of $\mathcal{J}\&\mathcal{F}$ performance decrease. Furthermore, the omission of instance trajectory generation via Hungarian algorithm [20] and temporal enhancement in Eq. (6), as well as trajectory ensemble in Eq. (7) (4^{th} row), result in a performance drop of 2.5% of $\mathcal{J}\&\mathcal{F}$. The empirical evidence suggests that importance of each component.

4.5. Qualitative Results

As shown in Fig. 6, we present the qualitative comparison between our approach and DsHmp [13] on the more challenging dataset, MeViS [7]. Specifically, the target (cat) in Fig. 6 (a) suffers from the distraction of the similar object. Except for the referring target, DsHmp erroneously segments the stationary cat as well. In contrast, SSA can segment the target precisely. For multi-instance referring scenario (Fig. 6 (b)), SSA accurately segments three riders on the elephants during the race. These qualitative results show that the proposed semantic and sequential alignment help the model better understand text-described objects and ensure consistent temporal segmentation.

5. Conclusion

In this paper, we propose a novel Semantic and Sequential Alignment framework for RVOS in real-world scenarios. Specifically, we achieve multi-modality semantic alignment by utilizing pretrianed VLM with proposed adapter for better video-to-text correspondence. Besides, the trajectory-to-instance correspondence established in the sequential alignment module enhances the model's ability to achieve global instance awareness. Extensive experiments on the challenging RVOS benchmarks demonstrate the effectiveness and generalization capacity of our method.

Acknowledgement

This work was supported in part by the Shandong Excellent Young Scientists Fund (ZR2024YQ006), Shandong Province Higher Education Institutions Youth Entrepreneurship and Technology Support Program (2023KJ027) and Taishan Scholars Program (Grant No. tsqn202312067).

References

[1] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeruIPS*, 2024. 3, 6

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *NeurIPS*, 2022. 3

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022. 1

[4] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 5

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 6

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[8] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *CVPR*, 2022. 6, 7

[9] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *ECCV*, 2024. 2, 3

[10] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 1, 2, 5

[11] 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. 3

[12] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *ICCV*, 2023. 2, 3, 4, 5, 7

[13] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024. 2, 3, 5, 6, 7, 8

[14] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *NeurIPS*, 2022. 2, 5

[15] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *AAAI*, 2025. 3, 6

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[17] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *NeurIPS*, 2023. 4

[18] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018. 1, 2, 5, 6, 7

[19] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip's image-text alignment to referring image segmentation. In *NAACL-HLT*, 2024. 3

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 5, 8

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

[23] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 3

[24] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 2

[25] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *ICCV*, 2023. 7

[26] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2024. 1

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3, 7

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 7

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 2

[33] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *NeurIPS*, 2023. 5, 7

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 6

[35] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *ICCV*, 2023. 7

[36] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6

[37] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *ECCV*, 2022. 3

[38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 5

[39] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[40] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1

[41] Zheyun Qin, Cheng Han, Qifan Wang, Xiushan Nie, Yilong Yin, and Lu Xiankai. Unified 3d segmenter as prototypical classifiers. *NeurIPS*, 2023. 1

[42] 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 *ICLR*, 2021. 3, 4, 5

[43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 6

[44] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3, 6

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 3

[47] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 1, 2, 5, 6, 7

[48] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 3

[49] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *ICCV*, 2023. 2, 3, 7

[50] 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. 3

[51] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[52] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 3

[53] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, 2023. 2, 3, 4, 5, 7

[54] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7

[55] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *ECCV*, 2024. 3, 6

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

[57] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[58] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 6

[59] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 3

[60] Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *CVPR*, 2024. 2, 7

[61] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 3