

QST-SAM: Leveraging Cross-modal Instructions for Few-shot Referring Video Object Segmentation

Xin Liu¹, Weijia Li¹, Tao Chen², Hongsong Wang³, Guo-Sen Xie⁴,
Caifeng Shan¹, and Fang Zhao^{1*}

¹ Nanjing University, Nanjing, China

² China Mobile Communications Group Jiangsu Co., Ltd., Nanjing, China

³ Southeast University, Nanjing, China

⁴ Nanjing University of Science and Technology, Nanjing, China

imxinliu@gmail.com, 602024710006@smail.nju.edu.cn,
chentao@js.chinamobile.com, hongsongwang@seu.edu.cn, gsxiehm@gmail.com,
caifeng.shan@gmail.com, zhaofang0627@gmail.com

Abstract. The advancement of Video Object Segmentation (VOS) is impeded by its heavy reliance on expensive annotations. This has led to a landscape of compromised alternatives: unsupervised methods provide weak information, few-shot learners are plagued by referential ambiguity from distractors, and referring techniques are limited by a lack of vision priors. We revisit Few-Shot Referring Video Object Segmentation (FSR-VOS) under a streamlined and practical formulation that segments target objects using only query-level textual descriptions and a small set of annotated support images. This dual conditioning enables coarse-to-fine disambiguation and precise boundary refinement with minimal annotation cost. To address the challenge of complex multi-source information fusion, we propose the QST-SAM framework. Specifically, a Support Instruction Compressor is introduced to refine and condense the support information into compact representations. The Q-S-T Transformer further facilitates comprehensive integration of multi-modal supervisory cues from query, support, and text branches. The Multi-Priors Prompting module provides the SAM2 prompt encoder with enriched guidance for precise, context-aware segmentation. We further establish a new FSRVOS benchmark derived from Ref-Youtube-VOS and Ref-DAVIS17, on which QST-SAM achieves state-of-the-art performance, validating its robustness and effectiveness in segmentation.

Keywords: Video Object Segmentation · Few-Shot Learning · Cross-modal Fusion

1 Introduction

The widespread use of video content has increased the demand for reliable Video Object Segmentation (VOS) [4, 5, 10, 20, 31, 39, 40, 47], which aims to segment

* Corresponding author.

and track target objects across frames. However, the prohibitive cost and labor-intensive nature of annotating video frames, combined with the rapid growth in video volume, high content diversity, and substantial domain shifts, make such approaches increasingly impractical for real-world deployment.

To reduce supervision, three major paradigms have emerged, each tackling a different aspect of the annotation burden. Unsupervised VOS [6, 7, 18, 25, 44, 52, 54] removes human input entirely, but without guidance, it often produces unstable or overly coarse masks. Referring VOS [9, 11, 32, 53] lets users specify the target using natural language, which improves flexibility. But when descriptions are vague or under-specified, models struggle to ground them to the correct visual instance. Moreover, the inherent gap between visual and textual representations amplifies this ambiguity, making cross-modal alignment particularly fragile in complex scenes. Few-shot VOS [3, 24, 29] provides a small number of example images to guide segmentation, reinforcing visual priors. Yet it falters when multiple similar-looking objects are present, or when the support and query frames differ too much in appearance, motion, or shape. In short, while these paradigms help reduce the reliance on dense annotations, they also introduce new sources of uncertainty.

This observation suggests a simple but powerful insight, that language and visual exemplars are complementary forms of supervision. Language encodes **intent** (what to segment), while support images encode **evidence** (what it looks like). Intent without evidence yields ambiguity, and evidence without intent yields mis-association. The remedy is to bind them. A recent effort [23] pioneered the **Few-Shot Referring Video Object Segmentation (FSRVOS)** task to unify the strengths of referring and few-shot paradigms. In their formulation, textual descriptions are provided for both the query video and each support episode. However, support-side descriptions are written for the corresponding support instances rather than the query video, and thus only share category-level semantics across videos, while other instance-specific details may not directly transfer. Such cross-video semantic mismatch can introduce redundant or less relevant signals when guiding query segmentation. In contrast, the query-level description already provides precise and context-aware intent for the target object. Therefore, we retain textual supervision exclusively at the query level and rely on masked support images to supply instance-level visual evidence, preserving complementary intent and appearance cues while reducing cross-video textual misalignment. Under this refined formulation, we study FSRVOS, which integrates referring expressions with few-shot visual exemplars under a streamlined supervision protocol. Given a query video and its textual description, together with a small set of masked support images from external sources, the goal is to segment the referred object consistently across the video. By combining query-level semantic intent with instance-level visual evidence from the supports, this setting enables precise and robust video object segmentation under reduced supervision.

However, while our refined formulation provides cleaner and complementary signals, it exposes a critical multi-modal alignment challenge: *how can language,*

support appearance, and spatiotemporal video cues be integrated in a way that helps, rather than hurts? Naive fusion is fragile, as textual intent and visual features differ in structure and granularity, support examples may carry spurious correlations, and temporal signals can propagate early mistakes. Effective FSRVOS requires, (1) filtering the support set to what truly grounds the target, (2) aligning this evidence to query frames at the instance level, and (3) guiding segmentation using stable, comprehensive priors that hold up even under occlusion, motion, or visual clutter. We introduce QST-SAM, a framework built on top of Segment Anything Model 2 (SAM2) [35], designed to reflect the structure of the FSRVOS problem. First, a Support Instruction Compressor distills the raw support set into compact, instruction-like representations—filtering out redundancy and spurious signals while preserving defining features of the target. Second, the Q-S-T Transformer integrates the distilled support cues with query features and text expressions, enabling the query video to jointly leverage visual evidence from the supports and semantic intent from the language, producing a robust and context-aware target representation. Finally, the Multi-Priors Prompting module fuses language and visual priors to condition the SAM2 prompt encoder, producing proposals that are semantically meaningful and visually grounded. In short, QST-SAM turns fragmented supervision into coherent, actionable guidance for segmentation. Furthermore, to evaluate the proposed FSRVOS setting, we establish new benchmarks derived from Ref-Youtube-VOS and Ref-DAVIS17. We then adapt state-of-the-art RVOS methods to this protocol, revealing that these strong baselines struggle, particularly when encountering semantic ambiguity or multiple similar instances. In contrast, QST-SAM exhibits clear advantages in fusing complex information, achieving new state-of-the-art performance.

In summary, our main contributions are as follows:

- We revisit the Few-Shot Referring Video Object Segmentation task and introduce a cleaner and more principled supervision protocol. By explicitly decoupling semantic intent from visual exemplars, our formulation reduces cross-video textual misalignment, avoids redundant supervision, and exposes the core multimodal grounding challenge in a more controlled and realistic setting.
- We propose QST-SAM, built upon *Segment Anything Model 2*, to integrate language, support appearance, and spatiotemporal cues. Through support distillation, cross-modal alignment, and multi-prior prompting, the framework produces robust and coherent segmentation under ambiguity and appearance variation.
- We establish new benchmarks based on *Ref-Youtube-VOS* and *Ref-DAVIS17*, and adapt strong RVOS baselines for comparison. Results show that existing methods struggle with semantic ambiguity and instance confusion, while QST-SAM achieves consistent state-of-the-art performance.

2 Related Work

Few-Shot Video Object Segmentation. Few-Shot Video Object Segmentation (FSVOS) [27, 48] aims to segment objects in query videos given only a few labeled support images of the same category. Existing methods either perform dense affinity matching between support and query frames [3] or summarize support features into class-specific prototypes to guide segmentation [24, 29]. While effective, these approaches rely solely on limited visual supervision, making it difficult to distinguish targets when multiple instances of the same class coexist or when visually similar but semantically different objects appear. Moreover, large intra-class variations in appearance, pose, or scale cannot be fully captured by a few support examples. Incorporating textual descriptions as auxiliary guidance provides explicit semantic cues, helping disambiguate targets and improving segmentation consistency across frames.

Referring Video Object Segmentation. Referring Video Object Segmentation (RVOS) [13, 19, 28, 41, 42] aims to segment objects in videos based on natural language expressions describing their identity, appearance, or actions. Early approaches [1, 12, 16] extended image-based referring segmentation to the video domain using recurrent architectures or spatio-temporal memory modules for mask propagation. Transformer-based methods [22, 43, 46], such as MTTR [2] and ReferFormer [43], further improved cross-modal alignment by jointly encoding textual and multi-frame visual features. More recent studies [14, 22, 45, 51] leverage pretrained multimodal models to aggregate visual-textual cues in a shared feature space before adapting them to the RVOS objective.

Despite these advances, RVOS still suffers from the inherent ambiguity of natural language, which often fails to uniquely identify the target instance in complex scenes. Existing attempts mitigate this issue through mechanisms such as memory refinement (SAMWISE [8]) or confidence-aware query pruning (ReferDINO [21]), yet the ambiguity problem remains.

Few-Shot Referring Video Object Segmentation. Few-Shot Referring Video Object Segmentation (FSRVOS) unifies referring expressions and few-shot visual supervision to segment objects in query videos. The original formulation [23] provides textual descriptions for both query and support episodes, enabling cross-modal grounding while reducing annotation costs.

Despite its conceptual appeal, this design entangles query-level intent with support-side textual descriptions written independently for support instances. Although these descriptions share category-level semantics, they often encode instance-specific attributes that may not transfer across videos, potentially introducing cross-video semantic misalignment. Consequently, textual supervision from the support branch may become partially redundant when guiding query segmentation, while also increasing annotation overhead.

Motivated by this observation, we revisit FSRVOS under a streamlined formulation that decouples intent and evidence. Specifically, we retain textual supervision exclusively at the query level to encode semantic intent, while relying

on masked support images to provide instance-level visual evidence. This refined protocol preserves the complementary nature of language and appearance cues, reduces cross-video textual mismatch, and offers a cleaner and more practical supervision scheme for robust video object segmentation.

3 Preliminaries

3.1 Segment Anything Model

The Segment Anything Model 2 (SAM2) extends the original Segment Anything Model (SAM) [17] to enable generic segmentation for both images and videos. The base architecture maintains SAM’s tripartite structure, comprising an Image/Video Encoder for extracting high-level features, a Prompt Encoder for converting user input into localization embeddings, and a Mask Decoder that integrates these features and embeddings to predict the final segmentation mask.

SAM2 adopts a MAE pre-trained Hiera image encoder [36] as its vision backbone, providing a hierarchical representation structure for multi-scale feature extraction. The Prompt Encoder transforms primarily visual prompts into embedding tokens that guide object localization. To process video streams effectively, SAM2 introduces a temporal memory pipeline. This module maintains a Memory Bank that stores historical segmentation features as memory embeddings, which are continuously updated by a Memory Encoder using a first-in–first-out (FIFO) strategy. The Memory Attention module then retrieves and integrates relevant memory embeddings to condition the current frame’s features, enabling temporally coherent representations. Finally, the Mask Decoder utilizes these memory-conditioned features together with the prompt tokens to generate the segmentation mask for the current frame.

3.2 Problem Definition

Given an input query video $V = \{I_f\}_{f=1}^F$ with F frames, a referring expression and a support episode, FSRVOS aims to predict a set of binary masks $M = \{M_f\}_{f=1}^F$, where $M_f \in \mathbb{R}^{H \times W}$ denotes the mask of the referred object at frame f . The referring expression provides a natural-language description that specifies the target object within the video. The support episode is composed of N image-mask pairs, which belong to the same class with the query object. Crucially, the target object class must be novel (unseen) during the model’s training phase.

4 Method

As illustrated in Figure 1, our proposed QST-SAM framework is composed of three key components: the Support Instruction Compressor (§4.1), the Q-S-T Transformer (§4.2), and the Multi-Priors Prompt Encoder (§4.3).

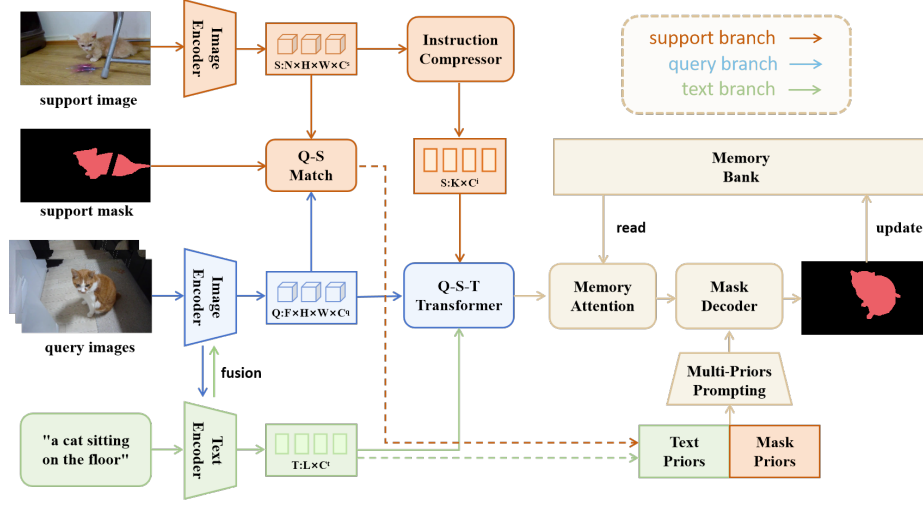


Fig. 1: Overview of our QST-SAM. QST-SAM integrates three branches: query, support, and text. Support features are compressed into compact support instructions via the Support Instruction Compressor. Query and text features are aligned through a cross-modal temporal adapter and then fused with the support instructions in the Q-S-T Transformer, which enables multi-directional interaction across query, support, and text. The Multi-Priors Prompting module establishes spatial correspondences between query and support features to derive mask priors, while transforming textual embeddings into complementary semantic priors. The multi-priors jointly guide the enhanced query feature towards robust mask prediction.

4.1 Support Instruction Compressor

Under the N -shot setting, the support branch first encodes support images using the frozen image encoder from SAM2, producing a high-dimensional feature map of support tokens $T_s \in \mathbb{R}^{N \times H_s \times W_s \times C}$. However, directly employing this dense feature map for cross-attention with query features leads to substantial computational overhead. Moreover, T_s inherently contains background noise and redundant information unrelated to the target object, which may corrupt the guidance signal and degrade segmentation performance.

To mitigate these issues, we introduce the **Support Instruction Compressor**, a lightweight module designed to distill T_s into a compact yet highly informative representation. The module first reduces the channel dimension of T_s via a linear projection, $\text{Linear}_{\text{down}}$. Subsequently, the down-projected features are processed by two parallel MLP branches, MLP_v and MLP_{attn} , to produce attention maps and content values, respectively, formulated as:

$$\begin{aligned} V_s &= \text{MLP}_v(\text{Linear}_{\text{down}}(T_s)), \\ \alpha_s &= \text{MLP}_{\text{attn}}(\text{Linear}_{\text{down}}(T_s)). \end{aligned} \quad (1)$$

Here, $V_s \in \mathbb{R}^{N \times H \times W \times C'}$ serves as the value matrix, encapsulating refined feature representations at each spatial location, where C' is the reduced channel dimension. $\alpha_s \in \mathbb{R}^{N \times H \times W \times K}$ denotes a set of K spatial attention maps, each capturing distinct and complementary regions or attributes of the support object to produce specialized instructions.

To suppress background noise, the learned attention maps α_s are refined by multiplying them with the interpolated support mask $\tilde{M}_s$, derived from the original mask M_s :

$$\alpha'_s = \sigma(\alpha_s) \odot \tilde{M}_s, \quad (2)$$

where $\sigma(\cdot)$ is a spatial sigmoid function used to normalize attention weights, and $\odot$ denotes element-wise multiplication.

Finally, the informative support instructions $\mathbf{S} = \{S_k\}_{k=1}^K$ are obtained through a weighted aggregation of value features V_s , guided by the refined attention maps α'_s :

$$S_k = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W \alpha'_s(n, i, j, k) V_s(n, i, j, :). \quad (3)$$

The resulting instruction set $\mathbf{S} \in \mathbb{R}^{K \times C'}$ effectively condenses the essential information from the support images into a compact representation that is resilient to background clutter and well-suited for interaction with query features.

4.2 Q-S-T Transformer

Overview. The Q-S-T Transformer serves as the central fusion module of our framework, enabling comprehensive interactions among the three distinct branches: query tokens $\mathbf{Q}$, support instruction tokens $\mathbf{S}$, and text tokens $\mathbf{T}$. As illustrated in Figure 2, each Q-S-T Transformer block iteratively merges linguistic semantics from the text expression, structural priors from the support episode, and spatial representations from the query. Specifically, each block is structured into two core sub-modules: the Text-Guided Semantic Refinement and the Spatio-Contextual Enhancement.

Feature Preparation and Alignment. Support instructions are obtained by compressing the support images using the Support Instruction Compressor described in Sec. 4.1. Query and text tokens are extracted using frozen image and text encoders and enhanced by a cross-modal temporal adapter, which facilitates the integration of images and text while incorporating temporal information. Note that the adapter is adopted directly from the SAMWISE framework without modification and its detailed implementation is provided in the supplementary material.

Text-Guided Semantic Refinement. Although support episodes correspond to the same target category, they often exhibit substantial intra-class variation in texture, viewpoint, shape, and motion. Such discrepancies introduce irrelevant

cues that weaken correspondence between query and support. To mitigate this, the initial stage of the Q-S-T Transformer performs text-guided semantic refinement on the support instructions $\mathbf{S}$ under the supervision of text tokens $\mathbf{T}$. This process is designed to emphasize support cues that are more relevant to the referring expression while reducing the influence of less relevant support information, so that the refined support representation can be better conditioned on the language description. To achieve this, we introduce a bidirectional interaction between the text and support modalities, consisting of **Text-to-Support (T2S)** and **Support-to-Text (S2T)** attention mechanisms:

$$\begin{aligned}\mathbf{T}' &= \text{FFN}\left(\text{SelfAttn}(\mathbf{T} + \text{CrossAttn}(\mathbf{T}, \mathbf{S}, \mathbf{S}))\right), \\ \mathbf{S}' &= \text{FFN}\left(\text{SelfAttn}(\mathbf{S} + \text{CrossAttn}(\mathbf{S}, \mathbf{T}', \mathbf{T}'))\right).\end{aligned}\tag{4}$$

Here, the T2S step allows text tokens to extract visual cues consistent with the linguistic description, yielding a support-aware textual representation $\mathbf{T}'$. The S2T step then propagates these refined semantics back to the support branch, enhancing features aligned with textual intent. Through this bidirectional refinement, the resulting $\mathbf{S}'$ is encouraged to be more closely conditioned on the linguistic description and less affected by irrelevant support cues, providing a more focused representation for subsequent query fusion.

Spatio-Contextual Enhancement. After obtaining refined support instructions $\mathbf{S}'$, we leverage them to guide the query tokens $\mathbf{Q}$ in capturing class-specific spatial and structural features. The second stage of the Q-S-T Transformer performs spatio-contextual enhancement through a bidirectional interaction between support and query modalities, implemented via **Support-to-Query (S2Q)** and **Query-to-Support (Q2S)** attention mechanisms:

$$\begin{aligned}\mathbf{S}'' &= \text{FFN}\left(\text{SelfAttn}(\mathbf{S}' + \text{CrossAttn}(\mathbf{S}', \mathbf{Q}, \mathbf{Q}))\right), \\ \mathbf{Q}' &= \text{Conv}\left(\text{SelfAttn}(\mathbf{Q} + \text{CrossAttn}(\mathbf{Q}, \mathbf{S}'', \mathbf{S}''))\right).\end{aligned}\tag{5}$$

In the S2Q step, the support instructions extract discriminative spatial cues such as category-consistent regions and motion patterns—from the query stream, producing a query-aware representation $\mathbf{S}''$. In the Q2S step, category-specific visual knowledge is propagated back to the query branch, enriching it with class-specific spatial organization and relational context. This reciprocal refinement encourages the final query features $\mathbf{Q}'$ to incorporate both query context and support-conditioned category cues, providing an enhanced representation for segmentation.

4.3 Multi-Priors Prompting Prediction

Traditional RVOS methods rely primarily on textual supervision, restricting the SAM2 prompt encoder to simple language prompts. In contrast, within the FS-RVOS setting, the support episode provides category-relevant visual information

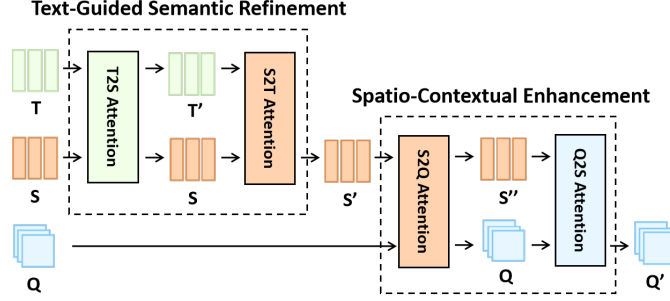


Fig. 2: Overview of the Q-S-T Transformer block. The Text-Guided Semantic Refinement module conditions support features on the text expression, while the Spatio-Contextual Enhancement updates the query stream with support-conditioned spatial and contextual cues.

that serves as strong visual priors. These priors effectively complement the inherent ambiguity of text-only descriptions, enriching spatial and appearance cues and thus improving the model’s ability to accurately localize and segment the referred object.

Q-S Matching Module. To establish reliable visual priors, the Q-S Matching module explicitly aligns query and support representations. Given a query feature map $T_q \in \mathbb{R}^{H_q \times W_q \times C}$ and N support feature maps $T_s \in \mathbb{R}^{N \times H_s \times W_s \times C}$, each support feature is first masked by its corresponding foreground annotation to suppress background interference. We then compute the affinity matrix $\mathbf{A}_n$ between the query and support tokens using cosine similarity, which measures the correspondence of feature embeddings across spatial positions.

To mitigate spurious correlations and focus on high-confidence affinities, we first sparsify the full affinity matrix $\mathbf{A}_n$. Instead of applying a fixed threshold, we adopt a simple percentile-based rule that computes a separate threshold for each matrix according to its own value distribution. This makes the filtering process adaptive and robust to variations between samples. Concretely, for each $\mathbf{A}_n$, we keep only the top $\tau\%$ of all affinity values and set the remaining ones to zero:

$$\hat{\mathbf{A}}_n(i, j) = \begin{cases} \mathbf{A}_n(i, j), & \text{if } \mathbf{A}_n(i, j) \geq \mathcal{T}_\tau(\mathbf{A}_n), \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\mathcal{T}_\tau(\mathbf{A}_n)$ is the τ -th percentile of $\mathbf{A}_n$. This step removes weak or irrelevant affinities while preserving the most confident spatial correspondences.

The filtered attention map $\hat{\mathbf{A}}_n$ is then used to propagate the structural prior from the n -th support mask $\tilde{\mathbf{M}}_s^n$ into the query domain. The final mask prior is obtained by averaging across all supports:

$$\mathbf{P}_m = \frac{1}{N} \sum_{n=1}^N \hat{\mathbf{A}}_n \tilde{\mathbf{M}}_s^n. \quad (7)$$

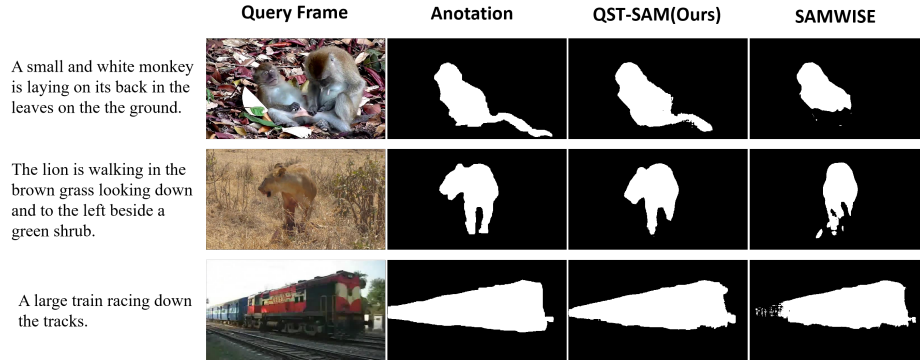


Fig. 3: Qualitative results on the Ref-YouTube-VOS dataset. By fully leveraging the visual prior knowledge derived from the support instruction, our QST-SAM demonstrates superior robustness against challenging conditions, such as heavy occlusion, low object-background contrast, and rapid motion.

In parallel to the visual prior $\mathbf{P}_m$, the text tokens are processed via an MLP projection to generate the text prior $\mathbf{P}_t$. The query feature $\mathbf{Q}'$ is enhanced by SAM2’s Memory Attention block, reading from the Memory Bank to yield the refined feature $\mathbf{Q}_{out}$. Finally, these three features are fused, where the decoder $\mathcal{D}_{dec}$ processes $\mathbf{Q}_{out}$ conditioned on $\mathbf{P}_m$ and $\mathbf{P}_t$ to generate the final segmentation mask $\mathbf{M}$:

$$\mathbf{M} = \mathcal{D}_{dec}(\mathbf{Q}_{out} | \mathbf{P}_m, \mathbf{P}_t). \quad (8)$$

5 Experiments

5.1 Datasets and Evaluation Metrics

We study FSRVOS, where a model is trained on a set of base categories and evaluated on novel categories using only a query text expression and N annotated support examples per episode. In this work, we focus on the 5-shot setting ($N = 5$) unless otherwise specified. As no standard benchmark exists for FSRVOS under our proposed support-text-free setting, we construct a new evaluation suite by reorganizing two established referring VOS datasets: Ref-YouTube-VOS [37] and Ref-DAVIS17 [16]. Furthermore, to ensure a fair and comprehensive comparison, we also evaluate our approach under the legacy FSRVOS setting [23].

Ref-YouTube-VOS. The original Ref-YouTube-VOS is a large-scale dataset for Referring Video Object Segmentation, comprising 3,978 videos and over 15,000 referring expressions. Annotations are provided only for its training split, which covers 3,471 videos across 65 distinct object categories. To enable few-shot evaluation, we partition these 65 categories into five disjoint folds of 13 categories each. We perform 5-fold cross-validation: in each fold, the model is trained on the 52 base categories (4 folds), and evaluated on the 13 held-out categories

(1 fold). To counteract the inherent long-tail distribution of data and ensure a stable evaluation, we sample 50 expression–video pairs per held-out category to form the validation set. The final score is the mean over all five folds, ensuring every category is evaluated exactly once.

Ref-DAVIS17. The Ref-DAVIS17 dataset extends DAVIS17 by incorporating natural language expressions, with its training split containing 60 videos. Following previous work, we train our model on the Few-Shot-Ref-YouTube-VOS dataset and directly test it on Ref-DAVIS without fine-tuning. To assess generalization to novel categories, we exclude the 25 overlap categories entirely from the Ref-YouTube-VOS training set (i.e., the model is trained on the remaining 40 base categories). During evaluation, each episode is constructed by sampling five random support image-mask pairs from held-out Ref-YouTube-VOS samples of the same category, with the corresponding Ref-DAVIS17 videos serving as the query set.

Evaluation Metrics. For quantitative evaluation, we utilize the standard metrics established by the DAVIS benchmark [33, 34]: Region Similarity ($\mathcal{J}$) and Contour Accuracy ($\mathcal{F}$). These metrics respectively gauge the quality of mask overlap, via Intersection-over-Union (IoU), and the precision of boundary alignment between our model’s predictions and the ground truth. To provide a comprehensive assessment, the primary reported score, denoted as $\mathcal{J}\&\mathcal{F}$, is the average of the $\mathcal{J}$ and $\mathcal{F}$ scores.

5.2 Implementation Details

For a fair comparison, our method follows the experimental setup of SAMWISE [8]. We employ Hiera-B [36] as the image encoder and RoBERTa [26] as the text encoder. The model is initialized with the SAMWISE weights pre-trained on RefCOCO+/g [30, 49]. We then fine-tune it on the Ref-YouTube-VOS dataset for 4 epochs, during which the text encoder and the entire SAM2 backbone remain frozen. The fine-tuning process uses the Adam optimizer with a learning rate of 1×10^{-5} and a video clip length of $T = 8$.

5.3 Comparison with State-of-the-Art Methods

To establish comprehensive baselines under our proposed decoupled FSRVOS protocol, we adapt state-of-the-art Referring Video Object Segmentation (RVOS) approaches to our newly partitioned datasets. For zero-shot evaluation, we retain their original architectures. For the few-shot scenario, we extend them by appending a lightweight fusion module that performs cross-attention between query and support features, followed by self-attention refinement (see supplementary material for architecture details).

Table 1: Comparison of zero-shot and few-shot performance on the Ref-DAVIS17 and Ref-YouTube-VOS datasets. The ‘(↑)’ column shows the improvement of the $\mathcal{J}\&\mathcal{F}$ score under the few-shot setting compared to the corresponding zero-shot baseline. Our proposed **QST-SAM** achieves the best overall performance and shows a significant gain over its baseline (SAMWISE).

Method	Venue	Setting	Ref-DAVIS17				Ref-YouTube-VOS			
			$\mathcal{J}\&\mathcal{F}$	(↑)	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	(↑)	$\mathcal{J}$	$\mathcal{F}$
DsHmp [15]	CVPR’24	Zero-shot	61.3	–	58.1	64.4	54.1	–	51.7	56.4
		Few-shot	64.5	(+3.2)	62.7	66.2	57.7	(+3.6)	56.4	58.9
LoSh [50]	CVPR’24	Zero-shot	60.4	–	57.4	63.4	55.9	–	53.2	58.5
		Few-shot	64.9	(+4.5)	62.2	67.5	58.1	(+2.2)	56.7	59.4
ReferDINO [21]	ICCV’25	Zero-shot	65.9	–	63.7	68.1	59.8	–	57.5	62.1
		Few-shot	67.8	(+1.9)	65.5	70.1	60.6	(+0.8)	58.4	62.8
SAMWISE [8]	CVPR’25	Zero-shot	65.5	–	63.5	67.4	59.3	–	57.2	61.3
		Few-shot	67.5	(+2.0)	65.3	69.6	60.4	(+1.1)	58.3	62.4
QST-SAM (ours)	–	Few-shot	70.3	(+4.8)	67.5	73.1	62.4	(+3.1)	60.9	63.9

Overall Performance. We evaluate our proposed QST-SAM against leading methods on the Ref-DAVIS17 and Ref-YouTube-VOS datasets. The comprehensive results, presented in Table 1, confirm that our approach sets a new state-of-the-art on both benchmarks. Specifically, on Ref-DAVIS17, QST-SAM achieves a remarkable 70.3% in J&F, surpassing the strongest competitor by a significant margin of 2.5%. On the large-scale Ref-YouTube-VOS benchmark, our method attains 62.4% J&F, outperforming the prior state-of-the-art by 1.8%. These consistent and substantial gains underscore the superiority of our framework in tackling the FSRVOS task.

Comparison under the New FSRVOS Setting. The core challenge in FSRVOS is to effectively integrate disparate guidance from text and support images, as naive fusion often proves insufficient to resolve visual ambiguities. Our results in Table 1 clearly validate this point. A simple adaptation of the powerful SAMWISE baseline yields only marginal J&F improvements (+2.0% on Ref-DAVIS17 and +1.1% on Ref-YouTube-VOS). Such limited gains reveal that the model struggles to translate the provided visual examples into effective segmentation cues. In stark contrast, our QST-SAM achieves a substantial leap of +4.8% and +3.1% over the same zero-shot baseline. This pronounced performance delta—more than doubling the gain on Ref-DAVIS17—directly validates that our proposed Q-S-T Transformer and Support Instruction Compressor are crucial. They successfully integrate the multi-modal instructions, unlocking a significant performance boost that a simple fusion approach cannot achieve and demonstrating a robust solution to the core FSRVOS problem.

Table 2: Quantitative performance comparison on the Mini-Ref-YouTube-VOS dataset. The Text_{sup} column indicates whether support text annotations are required. QST-SAM achieves state-of-the-art results across all metrics while relying on less support information.

Method	Text_{sup}	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
DANet [3]	✓	42.3	40.8	43.7
HPAN [38]	✓	50.9	51.1	50.7
CMA-ISM [23]	✓	54.0	53.1	54.8
QST-SAM (Ours)	×	59.2	58.2	60.1

Table 3: Ablation study of our key components on the Ref-YouTube-VOS dataset. The ‘ $\Delta\mathcal{J}\&\mathcal{F}$ ’ column shows the performance drop compared to the full model.

Configuration	$\mathcal{J}\&\mathcal{F}$	$\Delta\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
QST-SAM (Full Model)	62.4	—	60.9	63.9
w/o Q-S-T Transformer	60.8	(-1.6)	59.4	62.1
w/o Multi-Priors Prompting	61.7	(-0.7)	60.3	63.1
w/o both	60.4	(-2.0)	58.3	62.4

Comparison under the Legacy FSRVOS Setting. To ensure a fair and comprehensive comparison with prior FSRVOS methods, we also evaluate our approach on Mini-Ref-YouTube-VOS [23], which is a specific subset split of the Ref-YouTube-VOS dataset established under the legacy setting. As results summarized in Table 2, existing few-shot approaches rely on support-side textual annotations (Text_{sup}) to localize target objects. Despite exploiting this additional modality, the strongest prior method, CMA-ISM, achieves only 54.0% $\mathcal{J}\&\mathcal{F}$. In contrast, QST-SAM attains 59.2% $\mathcal{J}\&\mathcal{F}$ while entirely discarding support-level textual descriptions, surpassing CMA-ISM by +5.2% under strictly weaker supervision. This result indicates that discriminative visual cues derived from support masks can be more effective than cross-video textual annotations, which may introduce semantic inconsistencies.

These observations also support our revised FSRVOS formulation, suggesting that support-side textual descriptions provide limited additional information and may introduce redundancy in the supervision signal.

5.4 Ablation Studies

We conduct a series of ablation studies on the Ref-YouTube-VOS dataset to validate the efficacy of our key architectural components. Note that the Support Instruction Compressor and the Q-S-T Transformer are designed to be tightly coupled. Therefore, we ablate them as a single functional unit, treating the compressor as an integral component of the Q-S-T Transformer module. As

Table 4: Analysis of the number of support instructions. We evaluate the impact of varying the number of distilled support instructions (K) on the Ref-YouTube-VOS dataset, as well as the corresponding memory usage. Our final model uses $K = 32$.

Num. of Instructions	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Memory Usage (GB)
$K = 4$	59.8	58.1	61.4	38.1
$K = 8$	61.3	59.8	62.7	38.9
$K = 16$	61.8	60.4	63.2	39.7
$K = 32$ (Ours)	62.4	60.9	63.9	42.2
$K = 64$	62.6	61.4	63.8	47.6

detailed in Table 3, we systematically remove the Q-S-T Transformer and the Multi-Priors Prompting module to evaluate their impact on performance.

Q-S-T Transformer. Removing the Q-S-T Transformer causes the most significant performance degradation (-1.6 J&F), highlighting its role as the framework’s fusion core. Its effectiveness stems from a two-stage process: it first uses text to semantically refine support instructions—filtering visual noise—and then injects these purified structural priors into the query features. Without this structured fusion, the model fails to effectively align semantic guidance with visual evidence, leading to less accurate segmentation.

Multi-Priors Prompting. Ablating the Multi-Priors Prompting module results in a clear performance drop (-0.7 J&F). This demonstrates the value of providing SAM2 with an explicit visual mask prior alongside the text prompt. This visual prior, generated by our Q-S Matching module, provides a strong spatial localization signal that resolves the ambiguity inherent in text-only descriptions. Without this dual-modal guidance, the decoder lacks a precise spatial starting point, impairing its ability to generate accurate masks, particularly at object boundaries.

Details of Support Instruction Compressor. We analyze the effect of the number of support instructions K . As shown in Table 4, performance improves substantially as K increases from 4 to 32, peaking at **62.4 J&F**. This trend indicates that a sufficiently large instruction set is crucial for encoding a comprehensive and diverse set of the support object’s visual features. However, the performance gains saturate beyond this point; increasing K to 64 yields only a marginal improvement of +0.2 J&F while incurring additional computational cost. Therefore, we select $K = 32$ for our final model, as it achieves the best trade-off between performance and efficiency.

6 Conclusion

In this work, we revisit **FSRVOS** and propose a decoupled formulation that retains textual supervision only for query-level intent, effectively reducing an-

notation overhead and semantic mismatch. To leverage these multi-modal cues, we introduce **QST-SAM**, a SAM2-based framework that integrates compressed support instructions with linguistic semantics for precise segmentation. State-of-the-art results on Ref-YouTube-VOS and Ref-DAVIS17 demonstrate that our approach offers a robust and efficient solution for video object segmentation under limited supervision.

Acknowledgements

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM902), the National Natural Science Foundation of China (No. 62476124, 62302093, 62276134), the Natural Science Foundation of Jiangsu Province (No. BK20242015, BK20230833), and the Gusu Innovation and Entrepreneur Leading Talents (No. ZXL2025322).

References

1. Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J., Giro-i Nieto, X.: A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications* **82**(3), 4419–4438 (2023)
2. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4985–4995 (2022)
3. Chen, H., Wu, H., Zhao, N., Ren, S., He, S.: Delving deep into many-to-many attention for few-shot video object segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14040–14049 (2021)
4. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3151–3161 (2024)
5. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: *European conference on computer vision*. pp. 640–658. Springer (2022)
6. Cho, S., Lee, M., Lee, J., Kim, D., Lee, S.: Depthflow: Exploiting depth-flow structural correlations for unsupervised video object segmentation. *arXiv preprint arXiv:2507.19790* (2025)
7. Cho, S., Lee, M., Lee, S., Lee, D., Choi, H., Kim, I.J., Lee, S.: Dual prototype attention for unsupervised video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19238–19247 (2024)
8. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 3395–3405 (2025)
9. Deng, A., Chen, T., Yu, S., Yang, T., Spencer, L., Tian, Y., Mian, A.S., Bansal, M., Chen, C.: Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 8625–8636 (2025)

10. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268* (2024)
11. Fang, H., Cong, R., Lu, X., Zhou, X., Kwong, S., Zhang, W.: Decoupled motion expression video segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 13821–13831 (2025)
12. Gavriluk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5958–5966 (2018)
13. Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., Lu, H.: The devil is in temporal token: High quality video reasoning segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 29183–29192 (2025)
14. Han, S.H., Hyun, J., Lee, P., Shim, M., Wee, D., Kim, S.J.: Decomposed attention fusion in mllms for training-free video reasoning segmentation. *arXiv preprint arXiv:2510.19592* (2025)
15. He, S., Ding, H.: Decoupling static and hierarchical motion perception for referring video segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13332–13341 (2024)
16. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: *Asian conference on computer vision*. pp. 123–141. Springer (2018)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
18. Lee, M., Cho, S., Lee, D., Park, C., Lee, J., Lee, S.: Guided slot attention for unsupervised video object segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3807–3816 (2024)
19. Li, X., Wang, J., Xu, X., Yang, M., Yang, F., Zhao, Y., Singh, R., Raj, B.: Towards noise-tolerant speech-referring video object segmentation: Bridging speech and text. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pp. 2283–2296 (2023)
20. Li, X., Miao, D., He, Z., Wang, Y., Lu, H., Yang, M.H.: Learning spatial-semantic features for robust video object segmentation. In: *The Thirteenth International Conference on Learning Representations* (2024)
21. Liang, T., Lin, K.Y., Tan, C., Zhang, J., Zheng, W.S., Hu, J.F.: Referdino: Referring video object segmentation with visual grounding foundations. *arXiv preprint arXiv:2501.14607* (2025)
22. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: Glus: Global-local reasoning unified into a single large language model for video segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 8658–8667 (2025)
23. Liu, H., Li, G., Gao, M., Zhen, X., Zheng, F., Wang, Y.: Few-shot referring video single-and multi-object segmentation via cross-modal affinity with instance sequence matching. *International Journal of Computer Vision* **133**(8), 5610–5628 (2025)
24. Liu, N., Nan, K., Zhao, W., Liu, Y., Yao, X., Khan, S., Cholakkal, H., Anwer, R.M., Han, J., Khan, F.S.: Multi-grained temporal prototype learning for few-shot video object segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 18862–18871 (2023)
25. Liu, W., Shen, X., Li, H., Bi, X., Liu, B., Pun, C.M., Cun, X.: Depth-aware test-time training for zero-shot video object segmentation. In: *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19218–19227 (2024)
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
27. Luo, N., Wang, Y., Sun, R., Xiong, G., Zhang, T., Wu, F.: Exploring the better correlation for few-shot video object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024)
28. Luo, Z., Xiao, Y., Liu, Y., Li, S., Wang, Y., Tang, Y., Li, X., Yang, Y.: Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems* **36**, 26425–26437 (2023)
29. Mao, B., Liu, X., Shi, L., Yu, J., Li, F., Xiang, S.: Few-shot video object segmentation with prototype evolution. *Neural Computing and Applications* **36**(10), 5367–5382 (2024)
30. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: *European Conference on Computer Vision*. pp. 792–807. Springer (2016)
31. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9226–9235 (2019)
32. Pan, F., Fang, H., Li, F., Xu, Y., Li, Y., Benini, L., Lu, X.: Semantic and sequential alignment for referring video object segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 19067–19076 (2025)
33. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
34. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
35. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
36. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: *International conference on machine learning*. pp. 29441–29454. PMLR (2023)
37. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: *European conference on computer vision*. pp. 208–223. Springer (2020)
38. Tang, Y., Chen, T., Jiang, X., Yao, Y., Xie, G.S., Shen, H.T.: Holistic prototype attention network for few-shot video object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(8), 6699–6709 (2023)
39. Videnovic, J., Lukezic, A., Kristan, M.: A distractor-aware memory for visual object tracking with sam2. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 24255–24264 (2025)
40. Wang, H., Yan, W., Lai, Q., Geng, X.: Temporal consistency-aware text-to-motion generation. *Visual Intelligence* **4**(1), 7 (2026)
41. Wei, C., Zhong, Y., Tan, H., Zeng, Y., Liu, Y., Zhao, Z., Yang, Y.: Instructseg: Unifying instructed visual segmentation with multi-modal large language models. arXiv preprint arXiv:2412.14006 (2024)

42. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2761–2770 (2023)
43. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)
44. Xiangyu, Z., Songcheng, H., Wanyun, L., Xiaoqiang, L., Wei, Z.: Shallow features matter: Hierarchical memory with heterogeneous interaction for unsupervised video object segmentation. arXiv preprint arXiv:2507.22465 (2025)
45. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: European Conference on Computer Vision. pp. 98–115. Springer (2024)
46. Yang, Z., Wang, J., Ye, X., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
47. Yao, Y., Yang, Q., Cui, M., Bo, L.: Towards fine-grained interactive segmentation in images and videos. arXiv preprint arXiv:2502.09660 (2025)
48. Ying, K., Hu, H., Ding, H.: Move: Motion-guided few-shot video object segmentation. arXiv preprint arXiv:2507.22061 (2025)
49. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
50. Yuan, L., Shi, M., Yue, Z., Chen, Q.: Losh: Long-short text joint prediction network for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14001–14010 (2024)
51. Zhao, B., Wu, L.Y., Fan, X., Liu, D., Zhang, L., He, R., Shen, J., Li, X.: Unleashing hierarchical reasoning: An llm-driven framework for training-free referring video object segmentation. arXiv preprint arXiv:2509.05751 (2025)
52. Zheng, X., Li, W., He, S., Li, X., Zhang, W.: Intrinsic saliency guided trunk-collateral network for unsupervised video object segmentation. arXiv e-prints pp. arXiv-2504 (2025)
53. Zhu, Z., Feng, X., Chen, D., Yuan, J., Qiao, C., Hua, G.: Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In: European Conference on Computer Vision. pp. 452–469. Springer (2024)
54. Zhuge, Y., Gu, H., Zhang, L., Qi, J., Lu, H.: Learning motion and temporal cues for unsupervised video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems* **36**(5), 9084–9097 (2024)