remote sensing

MDPI

Article

Hyperspectral Video Object Segmentation Based on Spectral–Spatial Interaction Enhancement and Adaptive Correlation Memory Selection

Beibei Song [1], Yunfei Chen [1], Yuedong Tan [2] and Wenfang Sun [3,*]

[1] School of Information Engineering, Chang'an University, Xi'an 710064, China; bbsong@chd.edu.cn (B.S.); 2024124007@chd.edu.cn (Y.C.)
[2] Institute for Computer Science, Artificial Intelligence and Technology (INSAIT), Sofia University St. Kliment Ohridski, 1784 Sofia, Bulgaria; yuedong.tan@uni-wuerzburg.de
[3] School of Aerospace Science and Technology, Xidian University, Xi'an 710126, China
* Correspondence: wfsun@xidian.edu.cn

Highlights

What are the main findings?

- A hyperspectral video object segmentation dataset comprising hyperspectral videos and their corresponding false-color RGB videos derived from hyperspectral imaging (HSI) data is constructed. Frame-by-frame pixel-level annotations are provided for the visible (VIS), near-infrared (NIR), and red-edge near-infrared (RedNIR) subsets.
- Based on the SAM 2 framework, a spectral–spatial interaction enhancement mechanism is proposed. By integrating hyperspectral information with the spatial-structure features of the corresponding false-color RGB images, the representation of object details and target–background differences is enhanced.
- An adaptive correlation memory selection strategy is designed. According to inter-frame feature correlations, memory information is being dynamically screened and updated, thereby reducing the accumulation of unreliable information and improving memory reliability and temporal segmentation stability and target consistency in long video sequences.

What are the implications of the main findings?

- The proposed method improves segmentation accuracy, temporal stability, and overall robustness, and is particularly suitable for hyperspectral video scenarios with challenges such as occlusion, target deformation, similar-object interference, and complex backgrounds.

Abstract

Hyperspectral imaging (HSI) is an important imaging modality in the field of optical remote sensing, capable of simultaneously capturing spatial structures and continuous spectral responses related to material properties. Therefore, compared with conventional red-green-blue (RGB) images, it can provide richer representations for understanding dynamic scenes. However, most existing video object segmentation methods have been developed for RGB videos, while high-quality hyperspectral video datasets with frame-by-frame pixel-level annotations remain scarce. To address this gap, this study constructed a hyperspectral video object segmentation dataset, comprising hyperspectral videos and their corresponding three-channel false-color RGB videos derived from the same HSI data, with pixel-level annotations provided for each frame. The dataset contains three spectral subsets: visible (VIS), near-infrared (NIR), and red-edge near-infrared (RedNIR). Using this dataset,

Check for updates

Academic Editor: Salah Bourennane

Received: 11 October 2026
Revised: 31 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

we developed SSAM-SAM2, a hyperspectral video object segmentation network built on Segment Anything Model 2 (SAM 2). The network includes a spectral–spatial interaction enhancement (SSIE) module designed to identify informative spectral bands, suppress interference from redundant and noisy spectral channels, and integrate hyperspectral cues with the spatial-structure features of the corresponding false-color RGB images. In addition, the network introduces an Adaptive Correlation Memory Selection (ACMS) module that regulates memory sampling and selection according to variations in target state and inter-frame feature correlation. This design reduces the influence of redundant or unreliable historical information and improves temporal memory reliability under target deformation, motion variations, and background interference. Extensive experiments show that SSAM-SAM2 outperforms the selected comparison methods on the VIS, NIR, and RedNIR subsets under the evaluated settings. These results demonstrate the value of the proposed dataset and indicate that spectral–spatial enhancement and temporal adaptive memory mechanisms contribute to improving hyperspectral video object segmentation performance.

Keywords: video object segmentation; hyperspectral video; SAM 2; memory selection

1. Introduction

Video object segmentation is a fundamental yet highly challenging task in computer vision. It aims to accurately segment target objects throughout a video sequence while maintaining temporal consistency. Semi-supervised video object segmentation (SVOS) [1] predicts the pixel-level masks of target objects in subsequent frames, given their masks in the first frame of a video. Owing to its well-defined task setting and considerable practical value, SVOS has become an important research direction in the field of video object segmentation. As a fundamental technique for video content understanding, image and video object segmentation has been widely applied in medical and healthcare applications [2,3], agricultural and industrial inspection [4,5], autonomous driving [6], and satellite remote sensing [7,8], among other fields [9,10]. In recent years, rapid advances in deep learning have led to substantial progress in feature representation, temporal modeling, and target matching for video object segmentation, gradually improving the ability of existing methods to handle complex scenarios involving occlusion, deformation, and background interference [11]. Nevertheless, most existing studies focus on RGB videos, and their performance remains constrained by the limited range of information available from visible-light imagery.

Hyperspectral imaging is an important modality for optical remote sensing that simultaneously captures spatial details and spectral responses across multiple spectral bands. It has played an important role in meteorological forecasting, precision agriculture, geological exploration, environmental monitoring, and disaster observation [12,13]. Recent advances have further extended hyperspectral analysis to fine-grained target understanding. For example, twin-branch collaborative learning has been employed to capture complementary spatial and spectral representations for hyperspectral anomaly detection [14]. Advances in snapshot hyperspectral imaging have extended hyperspectral sensing from static-scene interpretation to dynamic video analysis. This development provides richer observations for continuous target perception in ground-based remote sensing, intelligent surveillance, and traffic monitoring [15,16]. In dynamic scenes, depth–area estimation has also been introduced into hyperspectral video tracking to improve target-state estimation and adaptation to substantial scale variation [17]. Compared with RGB images, which contain only three visible-light channels, hyperspectral images offer substantially higher spectral resolution and provide rich information about spectral characteristics and material properties. Such information enables more discriminative representations of objects that appear sim-

ilar but differ in material composition [16]. Introducing hyperspectral information into video object segmentation may therefore improve the distinction between targets and their backgrounds, while enhancing segmentation performance in challenging situations involving target deformation, partial occlusion, interference from similar objects, and complex backgrounds [18].

In video object segmentation, the difference between hyperspectral and conventional visible-light videos extends beyond the number of input channels. Conventional visible-light videos mainly represent objects through color, texture, and structural cues, which generally provide clear spatial details and are compatible with segmentation models pretrained on RGB data. However, these visual cues are sensitive to illumination changes, shadows, camouflage, and interference from objects with similar colors [15,16,18]. In contrast, hyperspectral videos record band-dependent responses over a wider spectral range and can reveal material-related differences that remain indistinguishable in visible-light images [12]. Such spectral information is particularly valuable when the target and background exhibit similar appearances but different spectral characteristics. Nevertheless, hyperspectral video segmentation also introduces additional challenges, including high-dimensional and redundant spectral information, band-dependent noise, reduced effective spatial resolution caused by snapshot spectral mosaics, and variations in the number and wavelength ranges of bands across different sensors [15,16]. Therefore, directly applying RGB-oriented video segmentation models to hyperspectral videos is insufficient. Effective hyperspectral video segmentation requires the joint consideration of informative-band selection, spectral redundancy suppression, spatial-detail preservation, and temporal target consistency.

The Segment Anything Model 2 (SAM 2) [19] is a general-purpose visual segmentation model designed for RGB images and videos. It propagates target information through a streaming memory mechanism and exhibits strong video processing capabilities. However, when target tracking and segmentation rely solely on RGB videos, the three visible-light channels often provide insufficient information for distinguishing the target from its surroundings. This limitation is particularly evident for objects that have similar colors but different material compositions. In such cases, the limited spectral information contained in RGB images cannot adequately characterize differences in material properties, and even well-performing visual tracking and segmentation models may produce suboptimal results because of the missing spectral cues [15]. Moreover, existing models generally employ encoding architectures designed for three-channel RGB images and therefore cannot be directly adapted to the high-dimensional spectral characteristics of hyperspectral data. Straightforward approaches to incorporating spectral data, such as spectral-band selection, channel concatenation, or dimensionality reduction, may result in the loss of important spectral information and prevent the distinctive advantages of hyperspectral data from being fully exploited. Consequently, extracting informative spectral features while suppressing spectral redundancy, and effectively integrating hyperspectral information with the spatial details provided by RGB images, constitute a key challenge in hyperspectral video object segmentation. The effectiveness of coordinating spatial structures with spectral discrepancies has also been demonstrated in hyperspectral anomaly detection through dual-stream spectral–spatial fusion [20]. Although anomaly detection differs from video object segmentation, this observation further supports the explicit joint modeling of spatial and spectral information in hyperspectral vision tasks.

In addition to multimodal feature representation, effective temporal memory management is also critical to the performance of video object segmentation. Most existing video segmentation models adopt memory storage strategies based on a fixed capacity, fixed update intervals, or recently processed frames, without adequately assessing the quality and usefulness of candidate memory frames [21]. In the presence of severe occlusion, rapid

target deformation, or complex background interference, low-quality predictions may be stored in the memory bank. The resulting errors can then propagate through subsequent frames, gradually leading to segmentation drift and mask degradation [22]. Hyperspectral videos involve more complex spectral–spatial–temporal relationships. If useful historical information cannot be selected dynamically according to the current target state, the noise introduced by a fixed memory strategy may further undermine segmentation stability [23]. Another limitation is the absence of hyperspectral video object segmentation datasets that provide both hyperspectral videos and their corresponding RGB representations with frame-by-frame pixel-level mask annotations. This shortage not only limits the availability of training samples but also hinders quantitative evaluation and comparison across different methods [24]. Hyperspectral video object segmentation therefore continues to face several challenges, including multimodal information fusion, effective temporal memory selection, and insufficient annotated data.

To address these challenges, we construct a hyperspectral video object segmentation dataset comprising hyperspectral videos and corresponding HSI-derived false-color RGB videos with pixel-level mask annotations for every frame. The dataset provides a basis for training, testing, and evaluating methods developed for this task. Building upon SAM 2, we further propose SSAM-SAM2, a hyperspectral video object segmentation network that integrates spectral–spatial interaction enhancement with adaptive correlation memory selection. Specifically, we design a Spectral–Spatial Interaction Enhancement (SSIE) module that adaptively selects informative spectral channels, suppresses spectral redundancy, and extracts discriminative hyperspectral features. Through spectral–spatial interaction, the module jointly enhances the spectral information of HSI and the spatial features extracted from the corresponding false-color RGB images. We also introduce an Adaptive Correlation Memory Selection (ACMS) mechanism, which dynamically evaluates and selects candidate temporal memories according to the correlation between current target features and historical memory features. By retaining highly reliable target information, ACMS reduces the interference caused by ineffective memories and the accumulation of erroneous predictions, thereby improving segmentation accuracy, temporal stability, and robustness in complex scenes. The main contributions of this work are summarized as follows:

1. A hyperspectral video object segmentation (HVOS) dataset is constructed, consisting of hyperspectral videos and their corresponding HSI-derived false-color RGB videos with frame-by-frame pixel-level mask annotations. The dataset alleviates the shortage of temporally annotated pixel-level data for this task and provides support for model training, quantitative evaluation, and method comparison.
2. We propose an SSIE module that adaptively selects informative spectral channels, suppresses spectral redundancy, and extracts discriminative spectral features. By enabling interactive fusion and joint enhancement of HSI spectral information and false-color RGB spatial information, the module allows the network to exploit the spectral discriminability of hyperspectral data more effectively, leading to improved target representation and recognition.
3. ACMS mechanism is introduced to improve the fixed-rule memory management strategy used in SAM 2. It dynamically selects reliable temporal memories based on correlations between target features, reducing error accumulation and segmentation drift caused by low-quality memories. This design enhances the model's robustness to complex background interference, target occlusion, and dynamic appearance changes.
4. We conduct systematic experiments on HVOS and compare SSAM-SAM2 with mainstream RGB-based video object segmentation models. The results demonstrate the effectiveness and performance advantages of the proposed method for hyperspectral video object segmentation.

2. Related Work

2.1. Recent Advances in Hyperspectral Image Processing

Attention mechanisms are increasingly used in hyperspectral image processing to emphasize informative spectral bands and spatial regions while suppressing redundant or noisy responses. Kang et al. developed a spectral–spatial double-branch attention network that separately refines spectral and spatial features before integrating them for hyperspectral image classification [25]. To improve the modeling of local dependencies, Arshad et al. introduced a spectral–spatial neighborhood attention Transformer with a channel-wise shift strategy [26]. More recently, Lu et al. combined convolutional and Transformer branches with spectral–spatial attention to aggregate complementary local and global representations [27]. Wang et al. further incorporated contextual multi-head self-attention into a spectral–spatial convolutional Transformer, enhancing the interaction between local contextual features and long-range dependencies [28]. These studies demonstrate that attention-based modeling can improve the selection and interaction of discriminative spectral–spatial information. However, most of these methods focus on static hyperspectral image classification and do not explicitly address temporal target consistency, memory reliability, or frame-by-frame pixel-level segmentation in hyperspectral videos. Recent hyperspectral video tracking studies have progressively shifted from directly adapting RGB trackers toward explicit spectral–spatial–temporal modeling. SpectralTrack combines spectral–spatial feature fusion with temporal memory enhancement to exploit both material information and historical target states [29]. SASU-Net introduces spectral adaptive aggregation weighting and scale updating to improve tracking robustness under spectral interference and target-scale variation [30]. In addition, recent unified tracking frameworks jointly model spectral, spatial, and temporal conditions to strengthen target representation across challenging video sequences [31]. These developments demonstrate that effective hyperspectral video analysis requires not only informative spectral representations but also adaptive modeling of temporal and target-state variations.

Hyperspectral anomaly detection has likewise evolved toward more structured modeling of spectral differences, background distributions, and spectral–spatial dependencies. SDENet employs variational mapping and Transformer-based representation learning to amplify subtle spectral discrepancies between anomalies and complex backgrounds [32]. DFBSNet explores complementary information through dual frequency-domain branches and an adaptive fusion-and-selection mechanism [33]. In noisy hyperspectral observations, JointNet jointly optimizes image denoising and anomaly detection to reduce the loss of weak anomaly information during noise suppression [34]. S3CRAD instead combines superpixel-guided background inpainting with spectral–spatial constrained representation to improve background reconstruction and anomaly discrimination [35]. Collectively, these studies show that recent hyperspectral anomaly detection increasingly relies on the coordinated exploitation of spectral differences, spatial context, frequency information, and background structure.

2.2. Semi-Supervised Video Object Segmentation

Depending on how target information is carried through a video sequence, existing semi-supervised VOS methods can broadly be grouped into three categories: online fine-tuning, temporal mask propagation, and feature matching.

Early studies on semi-supervised VOS mainly relied on online fine-tuning. OSVOS [1] pioneered the one-shot semi-supervised VOS framework, in which a pretrained segmentation network is fine-tuned at inference time using the target mask provided in the first frame. In this way, the network is adapted to the specific target appearing in the test video. OnAVOS [36] extended this idea by introducing an iterative online adaptation scheme,

allowing the model to be continually updated with information from subsequent frames and thus better accommodate changes in target appearance. Although online fine-tuning usually produces an accurate representation of the target, it requires several rounds of gradient-based optimization during testing. The associated computational cost limits its applicability to long videos and real-time scenarios. Another drawback is that the model may become overly dependent on the target appearance observed in the first frame. Target drift can therefore still occur when the object undergoes deformation, occlusion, or substantial scale variation.

Temporal mask propagation and feature matching have since become the two major directions in VOS research beyond online fine-tuning [37]. Mask propagation methods exploit spatial and temporal continuity between neighboring frames. A target mask or other target-related information from the preceding frame is propagated to the current frame and subsequently refined by a segmentation network. MaskTrack [38] takes the mask predicted in the previous frame together with the current image as input to estimate the current-frame mask. RGMP [39] uses a gated propagation mechanism to improve information transfer between the reference frame and the current frame. With the growing use of Transformers, AOT [40] models temporal associations among multiple objects and maintains more consistent segmentation in complex scenes. DeAOT [41] further separates target-agnostic visual features from target-specific identity information, making feature propagation more effective. While such methods are generally efficient at inference, their reliance on local temporal continuity remains a limitation. Under severe occlusion, rapid target motion, or abrupt scene changes, errors introduced during propagation may accumulate across frames and eventually cause localization failure.

Feature-matching methods follow a different strategy. They encode target features extracted from the first frame and selected historical frames into reference representations, which are then matched against the current frame at either the pixel or object level to locate the target region [42]. The incorporation of memory mechanisms has considerably improved the ability of these methods to capture long-term temporal dependencies [21]. A spatiotemporal feature memory bank allows the current frame to be matched with multiple historical observations, reducing the dependence on step-by-step propagation between adjacent frames. This approach offers greater stability in long videos and in cases where the target is occluded or undergoes appearance changes [43]. Research along this line has also provided an important basis for the memory architecture adopted by video segmentation models such as SAM 2 [19].

Drawing on advances in feature matching and memory-based modeling, this study investigates their application to hyperspectral video object segmentation, with particular emphasis on improving temporal segmentation under complex scene conditions.

2.3. Memory-Based Video Object Segmentation

Memory-based video object segmentation methods have become a major focus of semi-supervised VOS research in recent years, largely because of their ability to model long-term temporal dependencies, recover targets after occlusion, and generalize across different scenes. The Space-Time Memory Network (STM) [42] was the first to systematically introduce an external memory mechanism into semi-supervised VOS. It encodes image–mask features from the first and historical frames into a memory bank and retrieves target information through dense spatiotemporal matching, substantially improving the use of long-term reference information. STCN [43] further refined feature representation and matching. Its more effective memory reading and propagation mechanisms strengthen the distinction between the target and background while reducing the adverse effects of redundant historical features on segmentation. Designed for long videos, XMem [21]

employs a hierarchical memory architecture that jointly manages sensory, working, and long-term memories. This architecture preserves representative historical information while controlling storage overhead, thereby alleviating feature interference and reduced inference efficiency in very long sequences. Building on XMem, XMem++ [44] improves the construction and use of long-term memory. More representative reference information is incorporated to strengthen target representation and improve target retrieval under challenging conditions such as severe occlusion and rapid motion. Cutie [45] instead emphasizes the joint modeling of object-level representations and pixel-level features. It uses object queries to improve the separation of foreground targets from distracting background regions, reducing the influence of background noise on memory matching and target prediction while maintaining a favorable balance between accuracy and efficiency. RDE [46] introduces a recurrent dynamic embedding mechanism that continuously updates target embeddings according to inter-frame dynamics. The resulting representation adapts to changes in target appearance, scale, and state, partly mitigating the limited temporal adaptability and accumulated drift associated with fixed embeddings. JointFormer [47] combines the global association modeling capability of Transformers with joint spatiotemporal feature fusion. By modeling spatial details together with global temporal relationships, it accounts for both frame-level local segmentation accuracy and video-level temporal consistency, yielding strong overall performance in complex dynamic scenes.

Despite these advances, most existing memory management methods update historical information according to fixed temporal intervals, predefined memory capacities, or general feature similarity. They rarely assess the reliability of candidate memories specifically under complex scene conditions. When low-quality prediction features are repeatedly written to the memory bank, erroneous information may propagate to subsequent frames and eventually cause segmentation drift. To address this limitation, we introduce an Adaptive Correlation Memory Selection strategy into the memory-based framework. It is intended to improve the reliability of historical memories and the temporal stability of predictions in hyperspectral video object segmentation.

2.4. Segment Anything Model 2

Building on the progress made by memory-based VOS methods, general-purpose visual segmentation models have further advanced video object segmentation toward unified frameworks and large-scale pretraining. Among them, the SAM 2 [19] combines general-purpose visual segmentation with a memory mechanism. It achieves high segmentation accuracy, temporal consistency, and zero-shot generalization in open-world scenarios, providing the foundation for our investigation of hyperspectral video object segmentation.

Benefiting from its general segmentation and temporal tracking capabilities, SAM 2 has been extended to a variety of application scenarios. YOLO-SAM2 [48] incorporates a YOLOv8 detector to generate segmentation prompts automatically, thereby reducing the need for manual interaction. SurgSAM2 [49] optimizes temporal modeling and inference efficiency for surgical videos to meet the stability and real-time requirements of medical applications. Det-SAM2 [50] improves automation and computational efficiency in long-video processing through automatic object prompt generation and optimized resource management. SAM2Long [51] employs multi-path inference and result selection to alleviate error accumulation over long sequences. DAM4SAM [22] introduces a distractor-aware memory design and an introspective update strategy to address ambiguities in target association caused by visually similar distractors in complex dynamic scenes.

With the development of hyperspectral video technology, SAM 2 has also begun to be applied to hyperspectral video analysis. HyperSAM [52] is built upon SAM 2 and stores historical target information in a memory bank. It selects historical frames relevant to the

current target state through correlation measurement, strengthening target association across frames. DeSAM2 [53] likewise develops a hyperspectral object tracking framework based on SAM 2. By introducing a detection-based self-prompting mechanism, it continually supplies the model with prior information about the target location and improves tracking continuity in dynamic environments. These studies provide preliminary evidence that SAM 2 can be applied to hyperspectral video tasks and also indicate the potential of integrating general-purpose segmentation models with hyperspectral information.

Despite substantial progress in semi-supervised VOS, most existing methods have been developed for RGB videos. Their feature extraction, target embedding, and memory matching processes rely primarily on visual cues from the three RGB channels, with limited consideration of the multiband structure and spectral–spatial coupling inherent in hyperspectral videos. Existing hyperspectral extensions of SAM 2 still exhibit insufficient spectral–spatial interaction, inadequate identification of informative spectral bands, and limited use of spectral features. Consequently, they cannot fully exploit the ability of hyperspectral data to distinguish targets with similar appearances but different material properties. Moreover, memory management in SAM 2 is mainly governed by fixed rules and does not adequately assess the relevance of historical memories to the current target state. Low-quality memory entries may therefore continue to interfere with subsequent predictions. To address these limitations, we develop SSAM-SAM2 for hyperspectral video object segmentation using SAM 2 as the base framework. The proposed model is intended to make more effective use of the spectral–spatial information in hyperspectral videos while improving the reliability of memory management and the stability of segmentation in complex temporal scenarios.

3. Methods

The overall architecture and data flow of the proposed SSAM-SAM2 network are illustrated in Figure 1. For each video frame, SSIE jointly processes the hyperspectral image and the corresponding three-channel false-color RGB image derived from it. SSIE selects informative spectral bands, refines the complementary information contained in the residual bands, and interacts the resulting spectral features with the spatial details of the false-color RGB image. The resulting three-channel spectral–spatial representation is subsequently fed into the SAM 2 framework. The Image Encoder extracts the current-frame features, which are integrated with reliable historical target information from the Memory Bank through Memory Attention. The Mask Decoder then predicts the target mask for the current frame. The predicted mask and the associated frame features are further used by ACMS to filter candidate memories, characterize the current target state, and adaptively control temporal memory selection and updating. The selected historical features are written into the Memory Bank and serve as temporal references for the subsequent frames. In this framework, SSIE enhances the spectral–spatial representation of the current frame, ACMS regulates the temporal information retained for future inference, and the basic SAM 2 architecture performs image encoding, memory interaction, and mask decoding. Section 3.1 briefly reviews the basic SAM 2 architecture, while Sections 3.2 and 3.3 describe SSIE and ACMS, respectively.

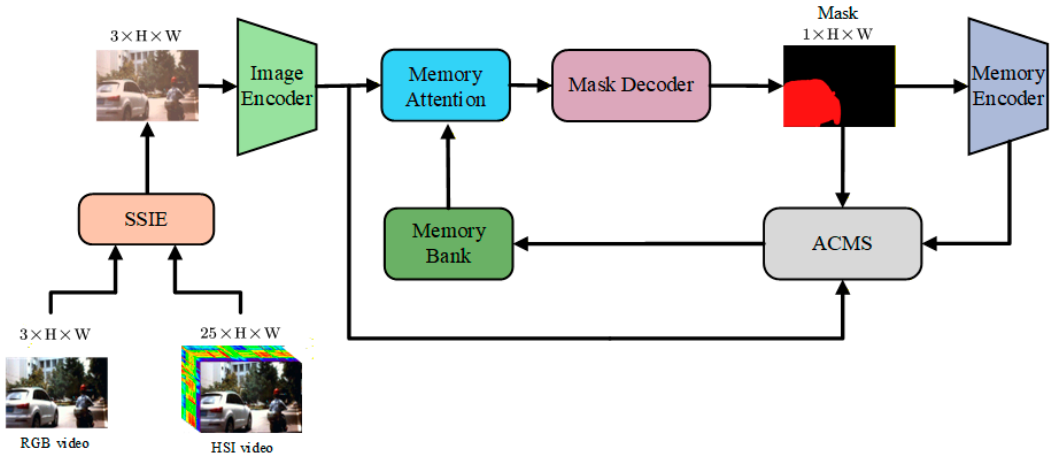


Figure 1. Overall architecture and data flow of SSAM-SAM2. For each frame, the hyperspectral image and its corresponding three-channel false-color RGB image are jointly processed by SSIE to generate a fused spectral–spatial representation compatible with SAM 2. The fused representation is encoded and combined with the historical information selected by ACMS through Memory Attention, after which the Mask Decoder predicts the current target mask. The predicted mask and frame features are subsequently used by ACMS for target-state characterization, candidate-memory selection, and Memory Bank updating, thereby supporting the segmentation of subsequent frames.

3.1. SAM 2 Preliminaries

The Image Encoder of SAM 2 employs Hiera [54] as its backbone to encode the input image. Hiera simplifies the conventional multiscale Vision Transformer architecture by removing redundant structures and certain specialized components, resulting in a more compact and efficient network. During pretraining, Hiera learns general visual representations using a Masked Autoencoder (MAE). Without relying on an elaborate multiscale design, it can effectively capture both shallow texture patterns and high-level semantic information. The resulting image embeddings support subsequent temporal modeling and mask prediction across the video sequence.

Memory Attention integrates the current-frame features produced by the Image Encoder with historical temporal information stored in the Memory Bank. The module is composed of multiple stacked Transformer layers. Self-attention is first applied to the current-frame features to model pixel-level relationships within the frame. Cross-attention then enables contextual interaction between the current-frame features and the spatial features and target semantics retained from historical frames. Through this process, the representation of the current frame can draw on previously observed target information, improving the model's ability to accommodate changes in target appearance.

Mask Decoder uses multiple self-attention and cross-attention layers to decode and fuse features of different dimensions. It produces several candidate segmentation masks, the corresponding predicted mask-IoU confidence scores, and an object-presence score for the current frame. Following a predefined candidate-selection rule, the mask with the highest confidence is taken as the final segmentation result for that frame.

Memory Bank stores and updates historical information using a First-In-First-Out (FIFO) queue. Its storage policy distinguishes between the initialization frame and recently processed frames. The supervised target features extracted from the initialization frame are retained permanently and serve as a global prior for subsequent segmentation. In parallel, target-tracking memory features from the six most recent frames are stored dynamically. Temporal positional encodings are assigned to these recent frames to indicate their chronological order. However, the original SAM 2 adopts a fixed-window memory update strategy and does not explicitly assess the quality of historical memories. During

continuous tracking, low-confidence or even erroneous target features may be written to the Memory Bank and subsequently used for feature matching. The resulting errors can accumulate from frame to frame, reducing segmentation accuracy and tracking stability in complex scenes.

3.2. Spectral–Spatial Interaction Enhancement Module

The original SAM 2 accepts only three-channel visual inputs and is not designed to process high-dimensional hyperspectral data directly. It therefore cannot fully exploit target-specific spectral differences, particularly when the target and background have similar spatial appearances but different spectral responses. To bridge this modality gap, we develop a Spectral–Spatial Interaction Enhancement (SSIE) module that transforms the hyperspectral data and the corresponding HSI-derived false-color RGB image into a three-channel spectral–spatial representation compatible with SAM 2.

SSIE draws on two groups of existing components: BandsGate and Band Embedding from HyperTrack [55], and Spectral Attention (SA) and Image-level Fusion (IF) from HyA-T [56]. We do not regard these individual operators as new contributions. Instead, the proposed modification lies in reorganizing them into a hierarchical band-routing, residual-compensation, and multi-branch fusion architecture specifically for SAM 2-based hyperspectral video object segmentation.

More specifically, BandsGate first ranks the aligned hyperspectral bands according to their learned importance. The highest-ranked bands are routed to a core spectral branch, whereas the remaining valid bands are retained in a separate residual-compensation branch. Unlike HyA-T, which directly performs attention and fusion using all hyperspectral bands, SSIE applies SA and IF only to the selected core bands. By restricting the interaction to the selected core bands, SSIE reduces interference from redundant or noise-dominated bands during fusion with the spatial features extracted from the false-color RGB image. Meanwhile, the residual bands are processed by Band Embedding to recover complementary spectral cues that may not receive high selection weights but can still contribute to target discrimination. The outputs of the core spectral–spatial interaction branch and the residual-compensation branch are subsequently combined with the false-color RGB spatial branch through learnable weighted fusion rather than direct channel concatenation. The resulting three-channel representation preserves the input compatibility of the pretrained SAM 2 Image Encoder. During training, SSIE and the SAM 2 backbone are optimized jointly using the segmentation objective, allowing the band-importance estimation, spectral refinement, residual compensation, and branch-fusion weights to adapt to the downstream segmentation task. Therefore, SSIE is not a sequential concatenation of existing components; it is a task-oriented reorganization that coordinates selective spectral enhancement, complementary information preservation, and modality adaptation within a unified front-end module.

The detailed structure of BandsGate is presented in Figure 2b. Its primary purpose is to select hyperspectral bands that provide strong target discrimination. Unlike conventional methods based on fixed band selection, BandsGate learns a target-saliency weight for each band adaptively through convolutional feature extraction and nonlinear mapping. The HVOS dataset contains three types of hyperspectral data with different numbers of bands: NIR with 25 bands, VIS with 16 bands, and RedNIR with 15 bands. To provide a consistent input dimensionality, we follow the band-alignment strategy proposed in our previous work [55]. Specifically, VIS and RedNIR data are zero-padded along the spectral dimension to 25 bands, allowing the three types of hyperspectral data to be processed in a standardized form.

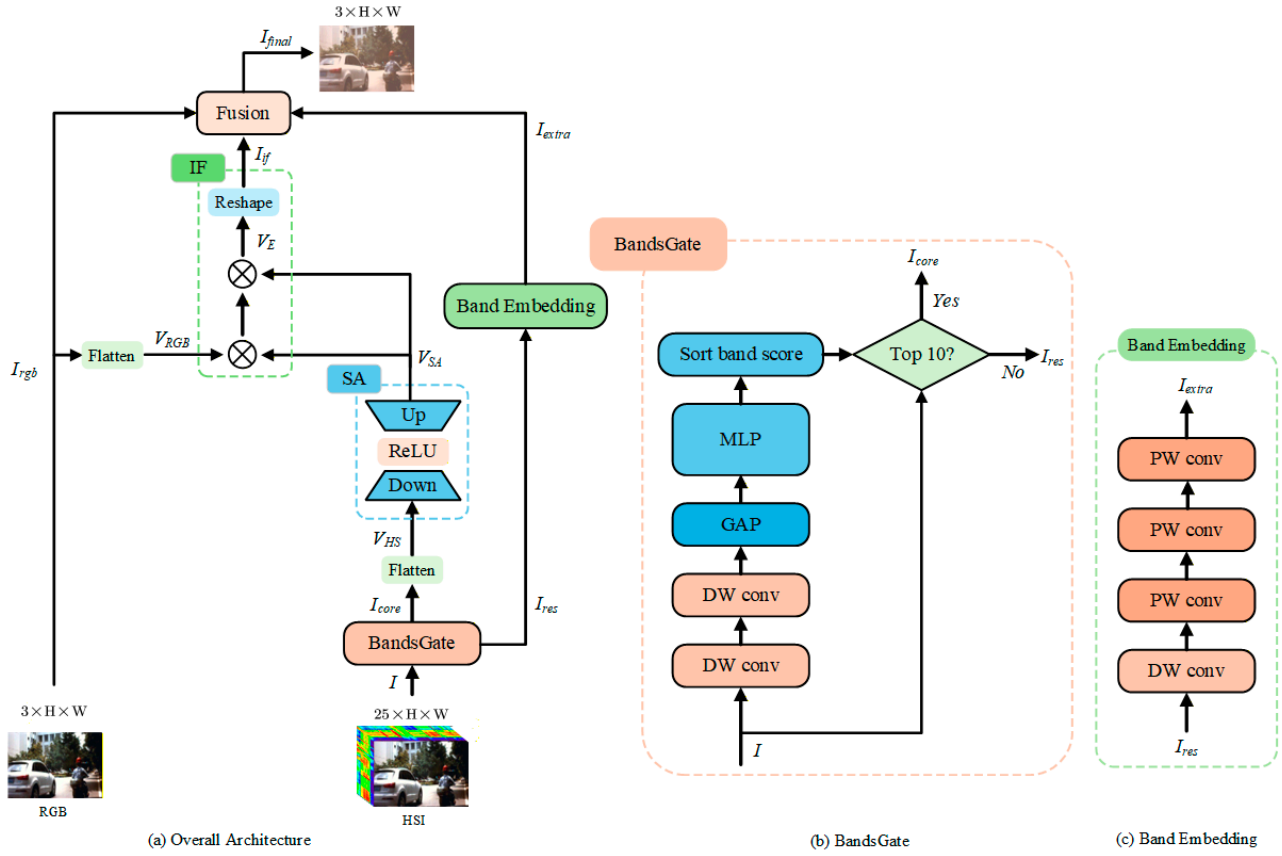


Figure 2. Architecture of the SSIE module: (**a**) Overall architecture; (**b**) BandsGate module; and (**c**) Band Embedding module. BandsGate routes the highest-ranked bands to the core spectral–spatial interaction branch, while Band Embedding transforms the remaining valid bands into residual-compensation features. The core interaction features, residual-compensation features, and false-color RGB spatial features are combined through learnable weighted fusion to generate the three-channel representation supplied to SAM 2.

More specifically, BandsGate consists of two depthwise convolution (DW) layers, global average pooling (GAP), and a multilayer perceptron (MLP). Two depthwise convolutions with kernel sizes of 7 × 7 and 3 × 3 are first applied to extract spatial features independently from each spectral band. This operation enlarges the spatial receptive field while preventing premature feature mixing across bands. GAP is then used to aggregate the feature map of each band over the spatial dimensions, producing a corresponding global statistical representation. Finally, these representations are passed through the MLP for nonlinear mapping to generate a saliency score for each band. This process can be formulated as follows:

$$S = \mathrm{MLP}(\mathrm{GAP}(\mathrm{DW}_{3\times3}(\mathrm{DW}_{7\times7}(I)))) \quad (1)$$

where $I \in \mathbb{R}^{25\times H\times W}$ denotes the hyperspectral input after zero-padding and band alignment, and $S \in \mathbb{R}^{25\times1\times1}$ denotes the saliency scores assigned to the individual spectral bands. The bands are ranked in descending order according to their saliency scores, and the ten highest-scoring bands are selected to form the core discriminative spectral branch, denoted by $I_{\mathrm{core}} \in \mathbb{R}^{10\times H\times W}$. The effectiveness of selecting the top ten bands has been experimentally validated in [55]. In addition to these core bands, the remaining bands $I_{\mathrm{res}} \in \mathbb{R}^{C_{res}\times H\times W}$ are retained to form a residual spectral compensation branch. This branch is intended to recover complementary discriminative information from bands with relatively weak responses. After excluding the zero-padded channels, the numbers of residual valid bands C_{res} are 15, 6, and 5 for the NIR, VIS, and RedNIR data, respectively.

The residual bands retained by BandsGate are processed independently by the Band Embedding module for feature embedding and dimensional transformation. Its detailed structure is presented in Figure 2c. The module comprises three pointwise convolution (PW) layers and one depthwise convolution (DW) layer and can be expressed as:

$$I_{\text{extra}} = \text{PW}(\text{PW}(\text{PW}(\text{DW}_{3\times3}(I_{\text{res}})))) \tag{2}$$

where I_{res} represents the residual-band features retained after BandsGate selection, and $I_{\text{extra}} \in \mathbb{R}^{3\times H\times W}$ represents the three-channel spectral compensation features produced by Band Embedding. The module further extracts and aggregates informative content from the residual bands, forming an independent spectral compensation branch. This branch complements the core discriminative spectral branch by preserving fine-grained spectral cues that may otherwise be overlooked. Through further exploitation of the potentially discriminative information carried by the residual bands, Band Embedding improves the utilization of hyperspectral data and strengthens the model's sensitivity to subtle spectral differences.

The SA and IF modules are employed to enable interaction and fusion between spectral and spatial information, as illustrated in Figure 2a. The SA module follows the low-rank projection attention architecture proposed in [56] and emphasizes informative spectral channels through adaptive weight assignment. In contrast to the original method, which directly takes all hyperspectral bands as input, our implementation feeds only the core-band features selected by BandsGate into the SA module. This modification directs the model toward spectral information with greater discriminative value. The computation is defined as follows:

$$V_{\text{HS}} = \text{Flatten}(I_{\text{core}}) \tag{3}$$

$$V_{\text{SA}} = W_{\text{up}}(\text{ReLU}(W_{\text{down}}(V_{\text{HS}}))) \tag{4}$$

where $V_{\text{HS}} \in \mathbb{R}^{10\times HW}$ denotes the flattened vector of the core hyperspectral bands; W_{down}, W_{up} denote the down-projection and up-projection operations in the SA module, and V_{SA} represents the refined spectral features produced by attention-based enhancement. These refined features are subsequently fed into the IF module, where they interact with the spatial features of the corresponding false-color RGB image through a parameter-free fusion operation. In this manner, material-discriminative information from the hyperspectral data is incorporated into the spatial texture representation of the corresponding false-color RGB image. The fusion process is defined as follows:

$$V_{\text{RGB}} = \text{Flatten}\left(I_{\text{rgb}}\right) \tag{5}$$

$$V_{\text{E}} = \text{Softmax}\left(V_{\text{RGB}} \times V_{\text{SA}}^{\text{T}}\right) \times V_{\text{SA}} \tag{6}$$

$$I_{\text{if}} = \text{Reshape}(V_{\text{E}}) \tag{7}$$

where $I_{\text{if}} \in \mathbb{R}^{3\times H\times W}$ denotes the three-channel spectral–spatial feature obtained through interaction between the core spectral bands and the false-color RGB features. This representation retains both spectral discriminability and fine-grained spatial structure. To integrate the complementary features produced by the different branches, we further design a learnable weighted fusion strategy. It aggregates the spatial details from the false-color RGB branch, the spectral–spatial features produced by the IF module, and the compensation features generated by Band Embedding. The fusion operation is formulated as:

$$I_{\text{final}} = I_{\text{rgb}} + w_1 \cdot I_{\text{if}} + w_2 \cdot I_{\text{extra}} \tag{8}$$

where $I_{\text{final}} \in \mathbb{R}^{3\times H\times W}$ denotes the final three-channel representation, and w_1, w_2 are learnable parameters used to balance the contributions of the false-color RGB spatial branch, the core spectral–spatial interaction branch, and the residual spectral-compensation branch.

The three branch outputs are combined by weighted summation rather than direct channel concatenation. These fusion parameters, together with the band-selection and feature-transformation components of SSIE, are optimized jointly with the SAM 2 backbone under the downstream segmentation objective. Consequently, the final representation is learned according to its contribution to target segmentation rather than being produced by a fixed combination of independently processed features. Accordingly, the quality of the representation delivered to SAM 2 does not depend exclusively on the core bands selected by BandsGate. If the selected core bands provide incomplete or locally unreliable cues, the residual-compensation branch retains information from all remaining valid bands, while the direct HSI-derived false-color RGB branch supplies stable target-boundary, texture, and spatial-structure information. The learnable weighted fusion further balances these complementary sources under the downstream segmentation objective, thereby reducing the sensitivity of the final representation to imperfect core-band selection.

Overall, the contribution of SSIE lies in the coordinated processing strategy rather than in the individual inherited operators. It first separates the hyperspectral bands into core and residual groups, performs attention-based spectral–spatial interaction only on the more informative core bands, and preserves complementary information from the residual bands through an independent compensation path. The two spectral branches are then adaptively balanced with the spatial representation of the HSI-derived false-color RGB image to produce a three-channel input for SAM 2. This hierarchical design reduces the interference of redundant bands without completely discarding weaker spectral responses and enables the resulting representation to be optimized jointly with the downstream segmentation network.

3.3. Adaptive Correlation Memory Selection Module

To clarify the relationship between ACMS and existing memory mechanisms, we distinguish its inherited foundations from the target-state-driven components introduced in this work. ACMS retains the memory organization of the Distractor-Aware Memory (DAM) framework in DAM4SAM [22] and adopts the inter-frame correlation measurement used in HyperSAM [52]. On top of these foundations, we introduce candidate-memory prefiltering, target-state statistics, adaptive sampling-stride control, and a dynamic correlation-selection window. Therefore, the contribution of ACMS lies not in redefining the basic memory branches or correlation operator, but in coordinating them through an adaptive selection strategy according to the current target state.

Inherited foundations. From DAM4SAM, ACMS retains the division of temporal memory into Recent Appearance Memory (RAM) and Distractor-Resolving Memory (DRM). DRM stores robust correction references, including the initialization frame and dynamically identified distractor frames, which provide a stable global target prior and help prevent tracking drift. DRM does not use temporal positional encoding and is updated only when the target is present, the required update interval has elapsed, and the tracking state is considered reliable. The key distractor frames are identified according to the disagreement among multiple mask hypotheses. RAM is responsible for short-term appearance modeling and retains the buffering mechanism, temporal positional encoding, and valid-target update constraint of the original framework. However, the original RAM is updated uniformly at a fixed interval of five frames and replaces memories primarily according to temporal order, without explicitly considering target scale or the degree of inter-frame variation.

ACMS also adopts the target-centered feature aggregation and depthwise grouped convolution-based correlation measurement from HyperSAM. This measurement estimates the similarity between historical and current target features and provides the correlation scores required for ranking candidate memories. Nevertheless, the inherited correlation

operator itself does not determine when correlation-based selection should be activated or how densely historical frames should be sampled under different target states.

Proposed adaptive components. On the basis of the inherited memory organization and correlation measurement, we introduce a target-state-driven memory selection strategy. Candidate frames with empty masks, frames already stored in DRM for distractor correction, and invalid frames without visual features are first excluded. The remaining valid historical masks are then used to calculate the mean target-pixel proportion and the coefficient of scale variation, which characterize the target scale and the degree of temporal deformation, respectively. Based on these statistics, two piecewise functions adapt the memory-sampling stride and correlation-selection window to the current target state. Within the dynamically determined correlation window, the inherited correlation scores are used to rank candidate frames, whereas outside this window, historical frames are sampled using the target-state-dependent stride. A temporally adjacent frame is additionally retained to preserve short-term continuity. In this way, ACMS converts the fixed memory update and correlation-selection process into an adaptive mechanism governed by target scale and temporal variation.

It should be clarified that ACMS is not a hyperspectral-specific operator in terms of its basic computation. It does not directly process spectral bands or wavelength-dependent responses, and its target-state statistics and inter-frame correlation logic can, in principle, be applied to other visual modalities. In the proposed framework, its HSI-VOS orientation arises at the system-integration level: ACMS operates on temporal features extracted by SAM 2 from the SSIE-enhanced spectral–spatial representation, and its target-state thresholds are determined according to the statistics of the HVOS training data. Therefore, the spectral information affects ACMS indirectly through the feature representation and correlation distribution, rather than through an explicitly spectral memory operator.

As illustrated in Figure 3, at intervals of $h_{\max}$ frames, the model examines the target states in the preceding $h_{\max}$ valid historical frames. A temporal sequence of target sizes is $\{n_1, n_2, \ldots, n_{h_{\max}}\}$ then constructed to calculate the mean target-pixel proportion mr and the coefficient of scale variation cv. These statistics are computed as follows:

$$\overline{n} = \frac{1}{h_{\max}} \sum_{i=1}^{h_{\max}} n_i \tag{9}$$

$$\sigma_n = \sqrt{\frac{1}{h_{\max}} \sum_{i=1}^{h_{\max}} (n_i - \overline{n})^2} \tag{10}$$

$$mr = \frac{\overline{n}}{H \cdot W} \tag{11}$$

$$cv = \frac{\sigma_n}{\overline{n}} \tag{12}$$

where n_i denotes the number of valid foreground pixels in the target mask of the i-th frame, obtained from the foreground region of the predicted mask. The symbol $h_{\max}$ is the length of the temporal statistics window and determines the number of valid historical frames used for each target-state estimate. The term $\overline{n}$ denotes the mean number of target pixels within this window, while σ_n is the population standard deviation of the target-pixel counts and characterizes the absolute magnitude of target-size fluctuations. The symbols H and W denote the height and width of the video image features, respectively. The mean target-pixel proportion, denoted by mr, describes the overall image scale of the target. The dimensionless coefficient of scale variation, denoted by cv, measures the relative temporal fluctuation in target size and reduces the influence of differences in absolute target size.

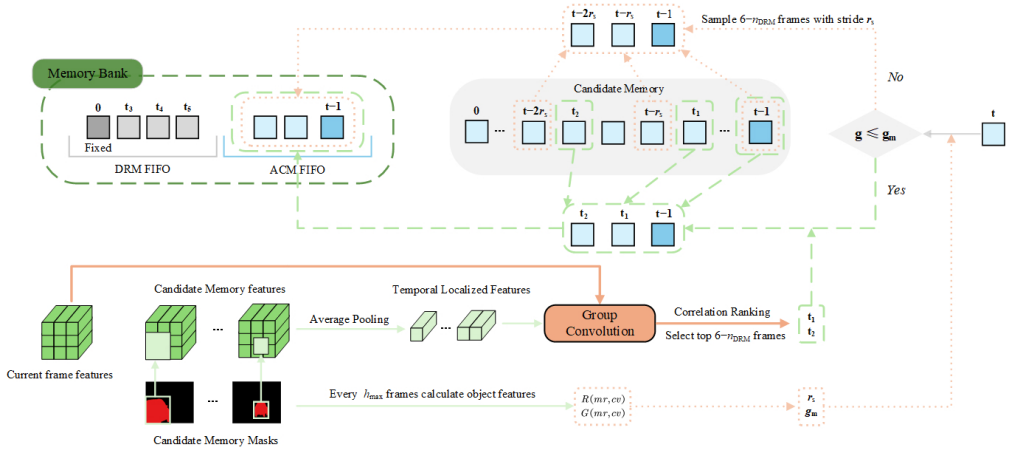


Figure 3. Architecture of the ACMS module. The memory organization inherited from DAM4SAM and the inter-frame correlation measurement adopted from HyperSAM provide the memory-storage and similarity-estimation foundations. Candidate-memory prefiltering, target-state statistics, adaptive sampling-stride control, and the dynamic correlation-selection window constitute the proposed components of ACMS. Solid arrows denote feature-processing and correlation-ranking flows, green dashed arrows denote memory sampling, selection, and update paths, and orange dotted arrows denote state-dependent control paths for the sampling stride and correlation window.

Unlike memory update strategies governed by fixed parameters, we define discrete piecewise control functions based on these two statistics. The functions switch the sampling stride and correlation-selection window adaptively according to the current target state:

$$R(mr,cv) = \begin{cases} R_{S,S}, & mr \leq mr_S,\ cv < cv_H \\ R_{S,V}, & mr \leq mr_S,\ cv \geq cv_H \\ R_{M,S}, & mr_S < mr \leq mr_M,\ cv < cv_H \\ R_{M,V}, & mr_S < mr \leq mr_M,\ cv \geq cv_H \\ R_{L,S}, & mr > mr_M,\ cv < cv_H \\ R_{L,V}, & mr > mr_M,\ cv \geq cv_H \end{cases} \tag{13}$$

$$G(mr,cv) = \begin{cases} G_{S,S}, & mr \leq mr_S,\ cv < cv_H \\ G_{S,V}, & mr \leq mr_S,\ cv \geq cv_H \\ G_{M,S}, & mr_S < mr \leq mr_M,\ cv < cv_H \\ G_{M,V}, & mr_S < mr \leq mr_M,\ cv \geq cv_H \\ G_{L,S}, & mr > mr_M,\ cv < cv_H \\ G_{L,V}, & mr > mr_M,\ cv \geq cv_H \end{cases} \tag{14}$$

where the subscripts S, M, L correspond to the small, medium, and large-target scale categories, respectively, whereas S, V indicate the Stable and Volatile states. The term mr_S denotes the scale threshold separating small and medium targets, and mr_M denotes the threshold separating medium and large targets. The term cv_H is the threshold used to distinguish different levels of target deformation fluctuation. The output R of the function $R(mr,cv)$ is used as the dynamic sampling stride *rank_stride* (r_s). It replaces the original fixed sampling interval of five frames and adapts the sparsity of memory updates to the current target state. The output G of $G(mr,cv)$ serves as the dynamic window threshold *max_gap* (g_m), which limits the effective temporal range for correlation-based selection. When target-state statistics are not computed, both functions use their default parameter settings.

The two target-scale thresholds, *mr_S* and *mr_M*, are determined from a statistical analysis of the target-scale distribution in the dataset. The results show that the mean target-pixel proportions are concentrated primarily between 1% and 10%. We therefore set the two thresholds to 0.02 and 0.10, respectively, to better reflect the distribution of target scales in the hyperspectral video scenes considered in this work. The setting of *cv_H* $h_{\max}$, and the specific outputs R and G of $R(mr, cv)$ and $G(mr, cv)$ under different combinations of scale and fluctuation states will be systematically investigated through controlled parameter experiments.

The proposed mechanism adaptively generates two core dynamic parameters according to the observed target state. For a small target undergoing substantial variation, a shorter sampling stride and a larger correlation window are used to increase the sampling density and capture transient appearance changes. By contrast, a larger sampling stride and a more restrictive correlation window are adopted for a large target with stable motion, reducing the update frequency and avoiding redundant memories. Compared with a fixed-parameter update strategy, this design is better able to accommodate the varying target conditions encountered in complex scenes.

The procedure for target-centered feature extraction and inter-frame correlation scoring is illustrated in Figure 3 and can be formulated as follows:

$$\phi_{x_i} = \mathrm{AvgPool}(\phi(x_i) \cap M), \quad i \in [t - t_s, t - 1] \tag{15}$$

$$\phi_{z_t} = \phi(z_t) \tag{16}$$

$$f(\phi_{z_t}, \phi_{x_i}) = \mathrm{Conv}(\phi_{z_t}, \phi_{x_i}, \mathrm{group} = C) \tag{17}$$

where t denotes the temporal position of the current search frame, and ϕ represents the image feature encoder. The encoded feature of the current frame is denoted by $\phi_{z_t} \in \mathbb{R}^{C \times H \times W}$ and serves as the input to the subsequent convolution operation. The term $M \in \mathbb{R}^{1 \times H \times W}$ denotes the downsampled and spatially aligned target-centered binary mask generated from the bounding box of the predicted mask. It is used to suppress irrelevant background features and retain features within the target region. The symbol t_s represents the temporal span over which historical candidate frames are selected. The target-centered feature of a historical candidate frame, denoted by $\phi_{x_i} \in \mathbb{R}^{C \times 1 \times 1}$, is obtained through mask-based cropping followed by mean pooling. During correlation computation, this feature is used as a channel-wise convolution kernel. Equation (17) performs inter-frame feature correlation matching through depthwise grouped convolution. The number of groups is set equal to the number of feature channels group $= C$. Each channel is therefore matched independently, preventing feature mixing across channels. The convolution produces a similarity score map over the spatial grid $\mathcal{D} \subset H \times W$. The peak response in the score map is taken as the final correlation score between the corresponding historical frame and the current frame. This score measures how closely the historical target appearance matches the current target state.

The proposed ACMS module first uses the output G of the piecewise function $G(mr, cv)$, i.e., g_m, as the threshold for determining the correlation-analysis window. If the temporal distance g between the current frame and the most recent DRM anchor frame does not exceed g_m, the target is considered to remain within the window in which correlation analysis is required. The anchor frame provides a fixed temporal reference throughout this window. Even if another frame is added to DRM before the window ends, the original DRM frame remains the anchor. This prevents accumulated errors that could otherwise arise from recursively replacing the temporal reference during correlation computation.

Within the correlation-analysis window, the module no longer uses conventional sampling with a fixed temporal stride. Instead, it computes correlation scores between the current-frame features and the candidate memory frames, ranks the candidates in

descending order, and selects the $7 - 1 - n_{DRM}$ highest-scoring frames. One memory slot is additionally reserved for a temporally adjacent reference frame to preserve continuity in the short-term temporal context. The n_{DRM} denotes the number of frames currently stored in DRM. The temporal selection span for historical candidate frames involved in correlation computation is set to 35 frames, providing sufficient temporal information while limiting the computational overhead. Once the target moves beyond this window and is no longer subject to the constraint imposed by the stable DRM anchor, the model reverts to interval-based temporal sampling. Candidate frames are then sampled using the dynamic stride R produced by the function $R(mr, cv)$, i.e., r_s. A temporally adjacent reference frame is also retained at this stage to maintain short-term continuity, resulting in a final selection of $7 - 1 - n_{DRM}$ candidate frames. Temporal positional encodings are assigned to the selected high-quality temporal memory frames. These frames are subsequently concatenated with the distractor-correction memories in DRM to form the final Memory Bank.

Overall, the RAM/DRM organization inherited from DAM4SAM and the correlation measurement adopted from HyperSAM provide the storage and feature-matching foundations of ACMS. The proposed contribution consists of candidate-memory prefiltering, target-scale and deformation statistics, adaptive control of the sampling stride and correlation window, and state-dependent switching between correlation-based selection and dynamic-stride sampling. Their integration enables memory samples to be selected according to the current target requirements, reducing redundant updates and limiting the propagation of unreliable historical information.

4. Experiments

This section describes the experimental setup, dataset construction, implementation details, and evaluation metrics. To evaluate the performance and effectiveness of SSAM-SAM2, we compare it with a range of video object segmentation methods. Ablation studies are further conducted to examine the contribution of each proposed module, while controlled parameter experiments are used to assess how different threshold combinations affect segmentation performance.

4.1. Datasets

Existing video object segmentation datasets, including DAVIS [57], YouTube-VOS [58], and MOSE [59], consist primarily of conventional RGB video frames and their corresponding segmentation masks. They do not provide detailed spectral information about the targets and are therefore insufficient for spectral–spatial feature discrimination and model evaluation in hyperspectral scenarios. To evaluate the proposed spectral–spatial enhancement and temporally adaptive memory mechanisms for hyperspectral video object segmentation, we construct a hyperspectral video object segmentation dataset containing paired HSI and HSI-derived false-color RGB representations, termed HVOS, based on the Hyperspectral Object Tracking (HOT) dataset [60]. HVOS covers three representative hyperspectral settings and provides the data required for investigating hyperspectral video object segmentation.

The HOT dataset was acquired using three XIMEA snapshot hyperspectral cameras covering the visible (VIS), near-infrared (NIR), and red-edge near-infrared (RedNIR) spectral ranges. Although the official HOT dataset description does not specify the exact commercial model of each camera, XIMEA xiSpec configurations matching the three spectral settings employ 4 × 4 VIS, 5 × 5 NIR, and 4 × 4 RedNIR snapshot spectral-filter mosaics, respectively [61]. These configurations are based on a 2.2-megapixel ams/CMOSIS CMV2000 CMOS detector with a native detector resolution of 2048 × 1088 pixels, a pixel size of 5.5 µm, a global shutter, and 10-bit digitization. Here, the native detector resolution refers to the resolution of the raw spectral mosaic rather than the effective spatial resolution of each spectral

band after mosaic reconstruction. Each HSI frame is reconstructed as a three-dimensional data cube consisting of two spatial dimensions and one spectral dimension. Accordingly, each spatial location contains spectral responses across multiple bands, enabling the data to characterize both spatial appearance and material-related spectral properties. For each HSI frame, the corresponding three-channel false-color image was generated from the same hyperspectral data using CIE color-matching functions; throughout this paper, these derived false-color images are referred to as false-color RGB images.

After calibration, the VIS, NIR, and RedNIR data contain 16, 25, and 15 spectral bands, respectively. The wavelength range is 460–620 nm for VIS, 625–975 nm for NIR, and 595–860 nm for RedNIR, collectively covering typical imaging wavelengths from the visible to the near-infrared spectrum. All videos were captured at a frame rate of 25 FPS.

When constructing HVOS, we retained the original partition assignment of every selected HOT video sequence. The VIS subset contains 55 training sequences and 46 test sequences; the NIR subset comprises 40 training sequences and 30 test sequences; and the RedNIR subset includes 15 training sequences and 11 test sequences. The split was performed at the complete video-sequence level, without randomly redistributing individual frames. All frames from a given sequence were therefore assigned exclusively to either the training set or the test set, preventing same-sequence frame leakage. No additional scene-level regrouping was performed beyond the original HOT partition.

Since the original annotations of HOT were intended primarily for object tracking, we further annotate pixel-level target segmentation masks using the corresponding false-color RGB images. The relatively clear spatial textures and target boundaries in these false-color RGB images facilitate the annotation of accurate binary foreground masks. For every frame, we also record the number of foreground pixels and the corresponding variation in target scale. These annotations and statistics provide supervision for hyperspectral video object segmentation and furnish empirical data for the subsequent adaptive parameter-control mechanism, ablation studies, and parameter sensitivity analysis.

4.2. Experimental Setup

4.2.1. Experimental Settings

The proposed model is implemented primarily using PyTorch 2.5.0 with CUDA 12.4 and trained in a distributed manner on two NVIDIA RTX 4090 GPUs. Some comparison models are trained on a single GPU. All SAM 2-based methods use sam2.1_hiera_base_plus as the base model. During training, the batch size is set to 1 per GPU, and eight frames are sampled from each video sequence. All input images are cropped to a resolution of 512 × 512, and the maximum number of training epochs is set to 120. AdamW is employed for optimization. The initial learning rate of the overall network is set to 5.0×10^{-6}, while that of the Image Encoder backbone is separately set to 3.0×10^{-6}. A cosine annealing schedule is adopted, with the final learning rate decaying to one-tenth of its initial value. To improve training stability and facilitate effective transfer of the pretrained model, layer-wise learning-rate decay is applied to the Image Encoder backbone with a decay factor of 0.9. Only the positional encoding parameters retain the global learning rate. Gradient clipping is also enabled during training, with the maximum gradient norm set to 0.1. Both the proposed SSIE module and the SAM 2 backbone are trained end to end. By contrast, ACMS is a training-free memory selection mechanism and introduces no additional learnable parameters; it is therefore excluded from gradient-based optimization.

4.2.2. Evaluation Metrics

Following the evaluation protocol of DAVIS [57], HVOS use region similarity, measured by the Jaccard index $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$. These metrics are calculated as follows:

$$J = \frac{M \cap G}{M \cup G} \tag{18}$$

$$F = \frac{2P_c R_c}{P_c + R_c} \tag{19}$$

$$J\&F = \frac{J + F}{2} \tag{20}$$

where M denotes the segmentation mask predicted by the model, and G denotes the ground-truth annotation. The symbols P_c and R_c represent contour precision and contour recall, respectively. The region similarity metric $\mathcal{J}$ measures the pixel-level overlap between the predicted and ground-truth masks and is a principal measure of region segmentation accuracy in video object segmentation. A higher $\mathcal{J}$ score indicates a more accurate prediction of the overall target region. Meanwhile, $\mathcal{F}$ focuses on the quality of the target boundary. It is typically computed by extracting single-pixel contours from the predicted and ground-truth masks through morphological operations, matching them within an adaptive boundary tolerance, and combining contour precision and contour recall. This metric reflects the model's ability to delineate target contours, fine structures, and edge details. It can therefore distinguish between segmentation results with similar region overlap but different levels of boundary accuracy. By combining region overlap and contour accuracy, $\mathcal{J}\&\mathcal{F}$ provides an overall assessment of model performance on hyperspectral video object segmentation.

4.3. Comparison Experiments and Analysis

To evaluate the effectiveness and generalization capability of SSAM-SAM2, experiments are conducted on the VIS, NIR, and RedNIR subsets of HVOS. The proposed method is trained and tested using hyperspectral video data. For a fair comparison, we select several representative RGB-based video object segmentation methods developed in recent years, including STCN [43], RDE [46], DeAOT [41], XMem [21], XMem++ [44], JointFormer [47], Cutie-B [45], SAM2-B+ [19], and the SAM 2 variants SAM2Long-B+ [51] and DAM4SAM-B+ [22]. All comparison methods are trained and evaluated separately using the HSI-derived false-color RGB videos from the corresponding subsets.

(1) Results on HVOS-VIS

The VIS subset contains 46 test video sequences, with each hyperspectral video comprising 16 spectral bands. Table 1 compares the proposed method with the RGB-based video object segmentation methods on HVOS-VIS. SSAM-SAM2 achieves the best performance across all three evaluation metrics, with $\mathcal{J}$ of 75.0%, $\mathcal{F}$ of 85.4%, and $\mathcal{J}\&\mathcal{F}$ of 80.2%. Compared with the second-best method, DAM4SAM-B+, the proposed method improves the $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ by 2.3, 2.2, and 2.2 percentage points, respectively. These results indicate that the proposed method effectively exploits both spatial texture and spectrally discriminative information in VIS hyperspectral data, improving the accuracy of target-region segmentation and boundary delineation.

(2) Results on HVOS-NIR

The NIR subset consists of 30 test video sequences, each containing 25 spectral bands. Table 2 reports the results of the different methods on HVOS-NIR. The proposed method achieves the highest scores on all three metrics, achieving $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ reaching 83.9%, 92.5%, and 88.2%, respectively. Specifically, the proposed method surpasses the second-best method for the $\mathcal{J}$, SAM2Long-B+, by 2.0 percentage points. For $\mathcal{F}$ and $\mathcal{J}\&\mathcal{F}$, it outperforms the respective runner-up, DAM4SAM-B+, by 1.1 and 1.6 percentage points. The results suggest that, under near-infrared imaging conditions, the proposed method can effectively

exploit multiband spectral information to distinguish the target from the background while maintaining stable segmentation over time.

Table 1. Comparison of different video object segmentation methods on HVOS-VIS. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	$\mathcal{J}$(%)	$\mathcal{F}$(%)	$\mathcal{J}\&\mathcal{F}$(%)	Params(M)	GFLOPs
STCN	2021	61.1	75.2	68.2	<u>54.43</u>	100.0
RDE	2022	45.7	54.8	50.3	64.02	120.9
XMem	2022	68.2	77.9	73.1	62.19	199.0
DeAOT-L	2022	61.5	73.6	67.5	70.30	135.1
XMem++	2023	66.4	76.0	71.2	62.19	184.9
Cutie-B	2024	61.1	71.7	66.4	**35.02**	**71.9**
SAM2-B+	2024	70.5	80.5	75.5	80.85	<u>91.8</u>
JointFormer	2025	65.2	73.0	69.1	127.67	468.5
SAM2Long-B+	2025	71.0	81.0	76.0	80.85	129.1
DAM4SAM-B+	2025	<u>72.7</u>	<u>83.2</u>	<u>78.0</u>	80.85	<u>91.8</u>
Ours	2026	**75.0**	**85.4**	**80.2**	80.86	92.3

Table 2. Comparison of different video object segmentation methods on HVOS-NIR. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	$\mathcal{J}$(%)	$\mathcal{F}$(%)	$\mathcal{J}\&\mathcal{F}$(%)	Params(M)	GFLOPs
STCN	2021	67.7	82.1	74.9	<u>54.43</u>	100.0
RDE	2022	73.3	84.1	78.7	64.02	120.9
XMem	2022	79.2	88.1	83.7	62.19	199.0
DeAOT-L	2022	76.6	86.5	81.5	70.30	135.1
XMem++	2023	79.2	88.0	83.6	62.19	184.9
Cutie-B	2024	74.3	85.4	79.9	**35.02**	**71.9**
SAM2-B+	2024	79.9	88.7	84.3	80.85	<u>91.8</u>
JointFormer	2025	78.2	84.9	81.6	127.67	468.5
SAM2Long-B+	2025	<u>81.9</u>	90.4	86.1	80.85	129.1
DAM4SAM-B+	2025	81.8	<u>91.4</u>	<u>86.6</u>	80.85	<u>91.8</u>
Ours	2026	**83.9**	**92.5**	**88.2**	80.86	92.7

(3) Results on HVOS-RedNIR

The RedNIR subset contains 11 test video sequences, with 15 spectral bands in each hyperspectral video. Table 3 presents the performance comparison on HVOS-RedNIR. SSAM-SAM2 obtains $\mathcal{J}$ of 59.2%, $\mathcal{F}$ of 67.6%, and $\mathcal{J}\&\mathcal{F}$ of 63.4%, outperforming all comparison methods on each metric. Relative to the second-best method, DAM4SAM-B+, the proposed method yields improvements of 0.7, 0.9, and 0.8 percentage points in the $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$, respectively. HVOS-RedNIR contains fewer test sequences and presents comparatively difficult segmentation conditions, including pronounced background interference, target-scale variation, and differences in spectral response. Nevertheless, SSAM-SAM2 delivers consistent improvements, further demonstrating its adaptability and robustness across different hyperspectral imaging ranges.

Overall, the proposed method achieves the best performance on the test sets of HVOS-VIS, HVOS-NIR, and HVOS-RedNIR. These results demonstrate the general applicability and effectiveness of the proposed spectral–spatial interaction enhancement mechanism and adaptive correlation memory selection strategy across different hyperspectral scenarios. SSAM-SAM2 makes effective use of the multiband information in hyperspectral videos, improving the discrimination between targets and complex backgrounds. Its more reliable temporal memory management also contributes to stable segmentation across consecutive frames.

Table 3. Comparison of different video object segmentation methods on HVOS-RedNIR. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	$\mathcal{J}$(%)	$\mathcal{F}$(%)	$\mathcal{J}\&\mathcal{F}$(%)	Params(M)	GFLOPs
STCN	2021	43.7	62.1	52.9	<u>54.43</u>	100.0
RDE	2022	39.7	48.9	44.3	64.02	120.9
XMem	2022	50.2	58.4	54.3	62.19	199.0
DeAOT-L	2022	47.4	64.4	55.9	70.30	135.1
XMem++	2023	50.2	58.5	54.4	62.19	184.9
Cutie-B	2024	45.9	54.9	50.4	**35.02**	**71.9**
SAM2-B+	2024	52.8	60.2	56.5	80.85	<u>91.8</u>
JointFormer	2025	49.3	57.3	53.3	127.67	468.5
SAM2Long-B+	2025	54.2	61.9	58.0	80.85	129.1
DAM4SAM-B+	2025	<u>58.5</u>	<u>66.7</u>	<u>62.6</u>	80.85	<u>91.8</u>
Ours	2026	**59.2**	**67.6**	**63.4**	80.86	92.2

Figures 4–7 present qualitative segmentation results for several representative video sequences. HVOS-VIS-paper (Figure 4) and HVOS-VIS-partylights6 (Figure 5) involve challenges such as background occlusion, illumination variation, and interference from visually similar objects. Under these conditions, most comparison methods either fail to recover the correct target shape or mistakenly segment a distractor, whereas SSAM-SAM2 maintains accurate target segmentation. The HVOS-NIR-pedestrian7 (Figure 6) and HVOS-RedNIR-rainystreet16 (Figure 7) sequences present additional difficulties, including small moving targets, target occlusion, and targets moving out of the field of view. Many of the comparison methods lose the target or switch to an incorrect object because of motion and occlusion, while SSAM-SAM2 handles these situations more reliably. Compared with the other methods, SSAM-SAM2 produces more complete and accurate target masks in complex scenes involving tiny moving targets, illumination changes, and occlusion, while preserving target boundaries and temporal consistency more effectively. These qualitative results further indicate that SSAM-SAM2 has strong adaptability to diverse hyperspectral scenes and promising potential for practical hyperspectral video object segmentation.

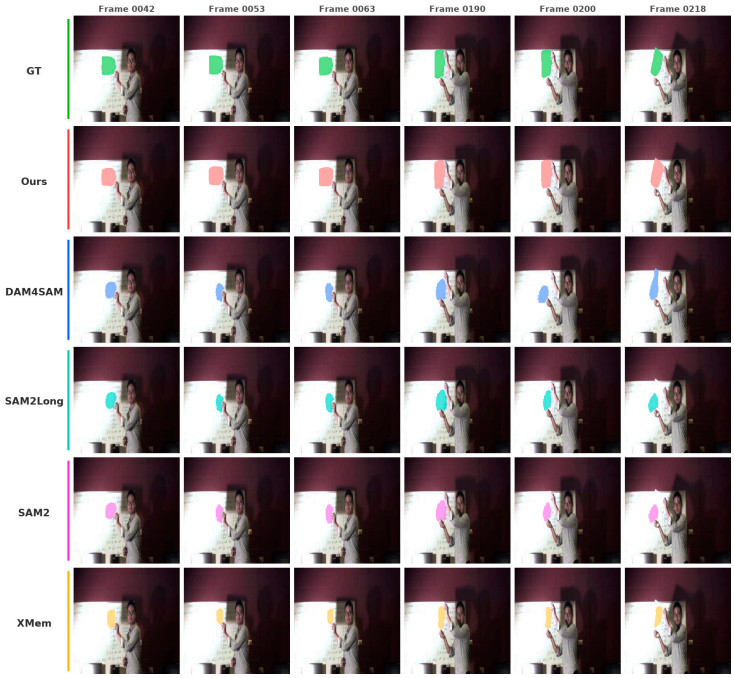


Figure 4. Qualitative comparison of segmentation results produced by different methods on the HVOS-VIS-paper sequence.

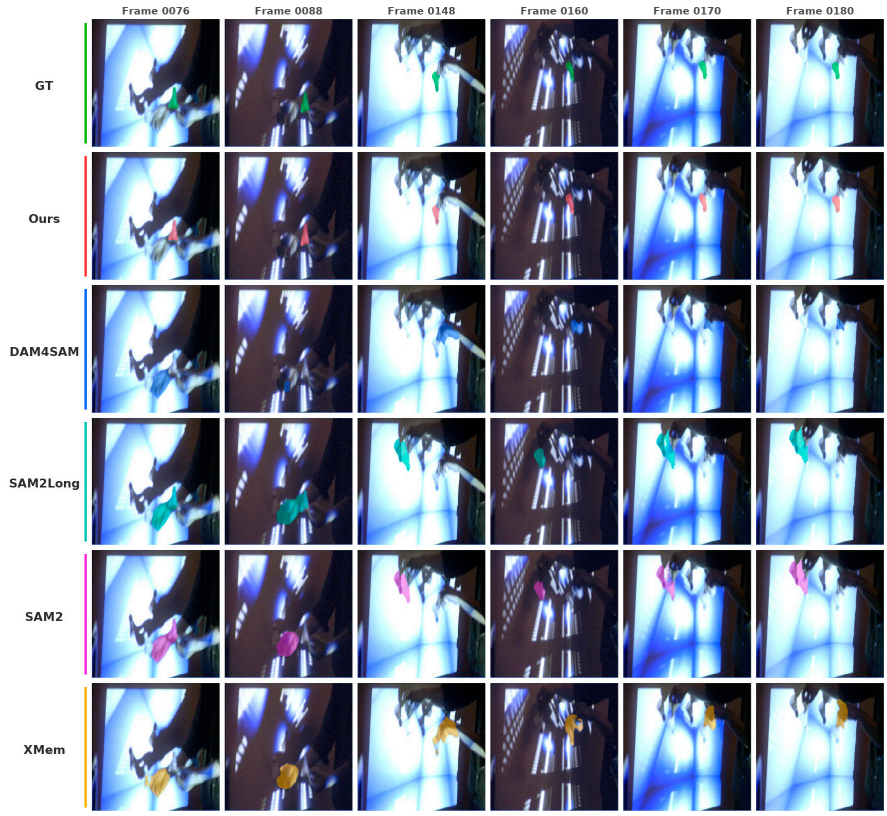


Figure 5. Qualitative comparison of segmentation results produced by different methods on the HVOS-VIS-partylights6 sequence.

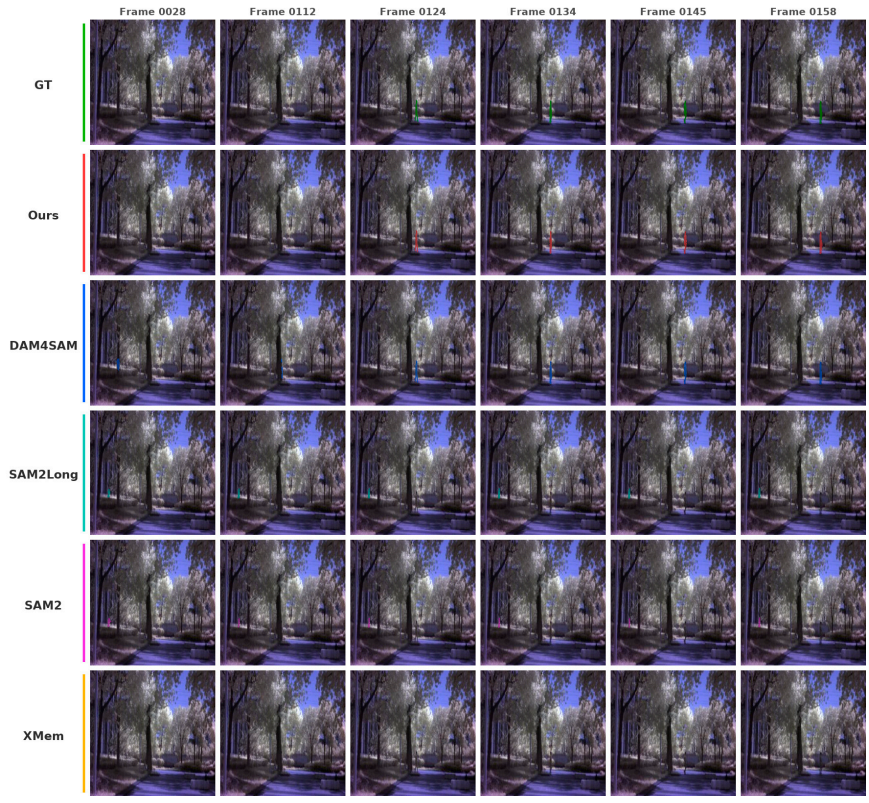


Figure 6. Qualitative comparison of segmentation results produced by different methods on the HVOS-NIR-pedestrian7 sequence.

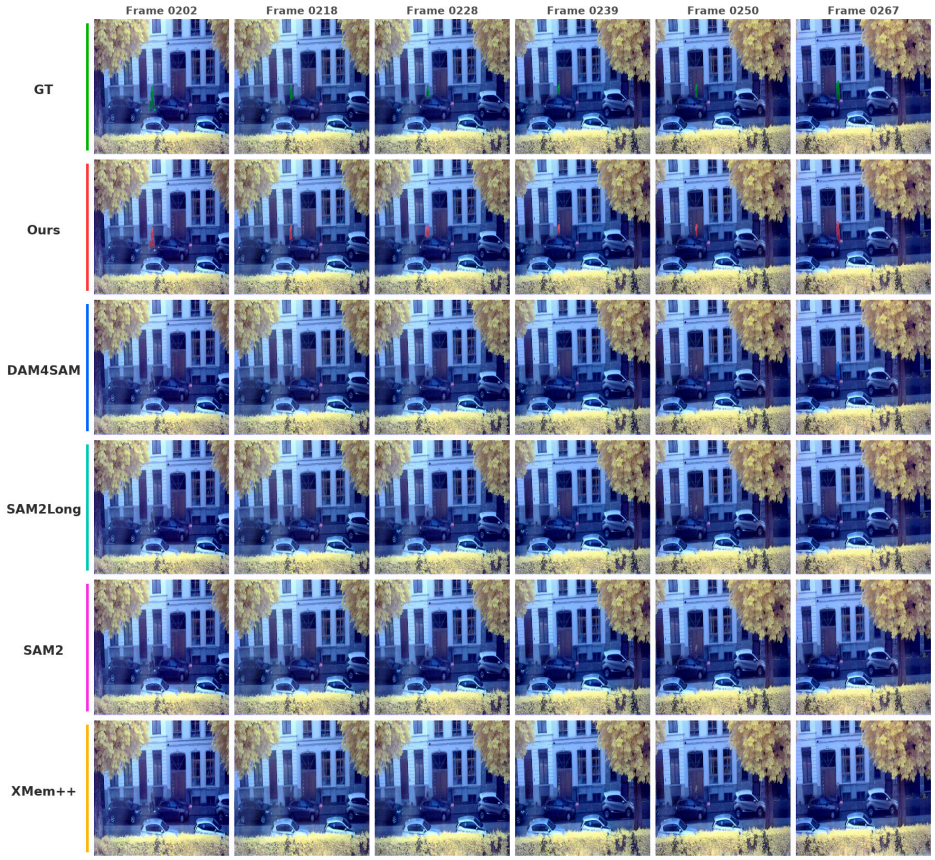


Figure 7. Qualitative comparison of segmentation results produced by different methods on the HVOS-RedNIR-rainystreet16 sequence.

4.4. Ablation Study

Ablation studies are conducted on all three subsets by selectively enabling the proposed components to assess their respective contributions to model performance. The results are reported in Table 4. The baseline is the original SAM2-B+ model, which obtains $\mathcal{J}\&\mathcal{F}$ of 75.5%, 84.3%, and 56.5% on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. The contribution of each subsequent component is analyzed relative to this baseline.

Table 4. Ablation results on the three HVOS subsets, including $\mathcal{J}\&\mathcal{F}$(%), Params (M) and GFLOPs. RGB denotes the HSI-derived false-color RGB input, and DFCB denotes the direct false-color branch. The best scores are shown in bold.

Type	Method	$\mathcal{J}\&\mathcal{F}$(%)			Params (M)	GFLOPs		
		VIS	NIR	RedNIR		VIS	NIR	RedNIR
RGB	Baseline	75.5	84.3	56.5	80.85	91.8	91.8	91.8
RGB	Baseline + ACMS	77.0	85.1	61.1	80.85	91.8	91.8	91.8
RGB	Baseline + DAM	78.0	86.6	62.6	80.85	91.8	91.8	91.8
RGB + HSI	Baseline + SSIE	78.1	84.9	57.7	80.86	92.3	92.7	92.2
RGB + HSI	Baseline + SSIE + DAM	79.1	83.7	59.9	80.86	92.3	92.7	92.2
RGB + HSI	Baseline + SSIE + ACMS	**80.2**	**88.2**	**63.4**	80.86	92.3	92.7	92.2
w/o DFCB	Baseline + SSIE + ACMS	51.0	22.3	33.9	80.86	92.3	92.7	92.2

(1) Effectiveness of SSIE

After integrating SSIE into SAM 2, the $\mathcal{J}\&\mathcal{F}$ scores increase by 2.6, 0.6, and 1.2 percentage points on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. These results show that SSIE compensates for the limited spectral modeling capability of the original RGB-

based video object segmentation model. It extracts discriminative features from multiband hyperspectral data while incorporating the spatial structure provided by the corresponding false-color RGB images, thereby improving the model's adaptability across different hyperspectral settings.

(2) Effectiveness of ACMS

When ACMS is further incorporated into the model equipped with SSIE, the $\mathcal{J}\&\mathcal{F}$ scores increase by a further 2.1, 3.3, and 5.7 percentage points on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. To examine the effectiveness of ACMS more thoroughly, we also replace it with the DAM mechanism from DAM4SAM while retaining SSIE. DAM improves $\mathcal{J}\&\mathcal{F}$ by 1.0 and 2.2 percentage points on HVOS-VIS and HVOS-RedNIR, respectively. On HVOS-NIR, however, DAM does not adequately utilize the hyperspectral information incorporated into the image features by SSIE. The segmentation performance of some targets consequently deteriorates, resulting in a 1.2-percentage-point decrease in $\mathcal{J}\&\mathcal{F}$.

Compared with DAM, ACMS achieves gains of 1.1, 4.5, and 3.5 percentage points in $\mathcal{J}\&\mathcal{F}$ on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. To examine whether this benefit depends exclusively on hyperspectral feature enhancement, we also apply ACMS to the false-color RGB-only Baseline. In this setting, ACMS improves the scores by 1.5, 0.8, and 4.6 percentage points, confirming that its target-state and correlation-based memory-selection logic is modality-agnostic. When ACMS is combined with SSIE, the corresponding improvements increase to 2.1, 3.3, and 5.7 percentage points. These results indicate that ACMS remains effective without explicit hyperspectral features, while its integration with the spectral–spatial representation produced by SSIE leads to larger gains on all three subsets. Therefore, the benefit of ACMS in the proposed framework should be attributed to system-level coordination between spectral–spatial representation and target-state-aware memory management, rather than to a hyperspectral-specific correlation operator.

(3) Complementarity of HSI, SSIE, and ACMS

To investigate the effectiveness and complementarity of hyperspectral data, SSIE, and ACMS, we additionally evaluate two false-color-RGB-only configurations: false-color RGB with ACMS and false-color RGB with DAM. Compared with the complete model combining HSI and false-color RGB inputs, SSIE, and ACMS, the RGB-only model equipped with ACMS exhibits decreases of 3.2, 3.1, and 2.3 percentage points in $\mathcal{J}\&\mathcal{F}$ on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. Meanwhile, compared with the RGB-only model equipped with DAM, the RGB-only ACMS configuration records $\mathcal{J}\&\mathcal{F}$ scores that are 1.0, 1.5, and 1.5 percentage points lower on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively.

These results further show that SSIE extracts discriminative information from HSI data at the feature level, thereby alleviating feature ambiguity in static scenes. When combined with SSIE, ACMS performs target-state-aware memory selection on the resulting spectral–spatial representation and improves temporal memory reliability, helping to reduce segmentation drift in dynamic hyperspectral scenes. SSIE and ACMS work collaboratively to form an integrated framework that combines spectral–spatial feature enhancement with temporally adaptive modeling. The experimental results also demonstrate the advantage of using full hyperspectral information over relying solely on its three-channel false-color RGB representation. Fine-grained spectral information provides material-discriminative cues that are difficult to represent using RGB images and thus forms an important basis for accurate hyperspectral video object segmentation.

To further investigate whether the complete HSI representation can compensate for the removal of the directly propagated false-color spatial features, we construct an additional variant by removing the direct false-color branch I_{rgb} from the final adaptive fusion. This variant is denoted as "w/o DFCB", where DFCB represents the direct false-color branch.

After its removal, the final SSIE output is generated only from the spectral–spatial interaction feature I_{if} and the residual spectral compensation feature I_{extra}:

$$I_{final} = w_1 \cdot I_{if} + w_2 \cdot I_{extra} \tag{21}$$

This experiment removes only the independently propagated false-color RGB pathway, not all false-color RGB information, because the IF-based spectral–spatial interaction retains the false-color features. This design allows us to evaluate whether the HSI-centered interaction and compensation branches can maintain segmentation performance without the direct preservation of the false-color RGB spatial representation.

As shown in Table 4, the w/o DFCB variant achieves $\mathcal{J}\&\mathcal{F}(\%)$ scores of 51.0%, 22.3%, and 33.9% on HVOS-VIS, HVOS-NIR, and HVOS-RedNIR, respectively. Compared with the complete Baseline+SSIE+ACMS model, the corresponding scores decrease by 29.2, 65.9, and 29.5 percentage points. These substantial reductions show that the spectral–spatial interaction feature and residual spectral compensation feature cannot independently preserve the target boundaries, textures, and stable spatial structures required by the pretrained SAM 2 backbone. Therefore, directly retaining the HSI-derived false-color RGB spatial representation remains important in the current transfer framework.

(4) Parameter Sensitivity of ACMS

ACMS contains several parameters that may affect segmentation performance. The parameter *cv_H* specifies the threshold used to determine the degree of inter-frame target-feature variation. The parameter h_{max} controls both the frequency of target-feature statistics and the maximum number of valid historical frames included in each statistical estimate. The output of the function $R(mr, cv)$, denoted as R, determines the sampling interval for candidate frames outside the non-correlation window, while the output of the function $G(mr, cv)$, denoted as G, defines the temporal extent of the correlation window.

To evaluate the influence of these parameters and identify an effective configuration, controlled experiments are conducted on four principal parameters of ACMS. The parameters are examined sequentially in the following order: *cv_H*, h_{max}, R, and G. All parameters are initialized to their default values. Once the best-performing value of a parameter is identified, it is fixed for the remaining experiments. The performance obtained with the resulting parameter configuration is presented in Figure 8. In each panel of Figure 8, the horizontal axis presents the candidate settings of the corresponding ACMS variable, while the vertical axis reports the mean segmentation score $\mathcal{J}\&\mathcal{F}(\%)$, with higher values indicating better segmentation performance.

The categorical labels on the R and G axes represent candidate output ranges of the two piecewise functions defined in Equations (13) and (14), rather than single scalar parameter values. Taking $R = (3 \sim 6)$ as an example, its candidate settings are $R_{S,V} = 3$, $R_{S,S} = R_{M,V} = 4$, $R_{M,S} = R_{L,V} = 5$, and $R_{L,S} = 6$. The candidate settings for $G = (3 \sim 6)$ are defined similarly as $G_{S,V} = 6$, $G_{S,S} = G_{M,V} = 5$, $G_{M,S} = G_{L,V} = 4$, and $G_{L,S} = 3$. The sampling-stride configurations follow the principle that larger targets are sampled more sparsely, whereas targets with greater fluctuations are sampled more densely. The correlation-window configurations follow the opposite trend: the window becomes smaller as the target scale increases and larger as target fluctuations become more pronounced. These settings allow the sampling frequency and correlation-window range to adapt to changes in target scale and temporal state, thereby improving memory-selection reliability under the evaluated video conditions.

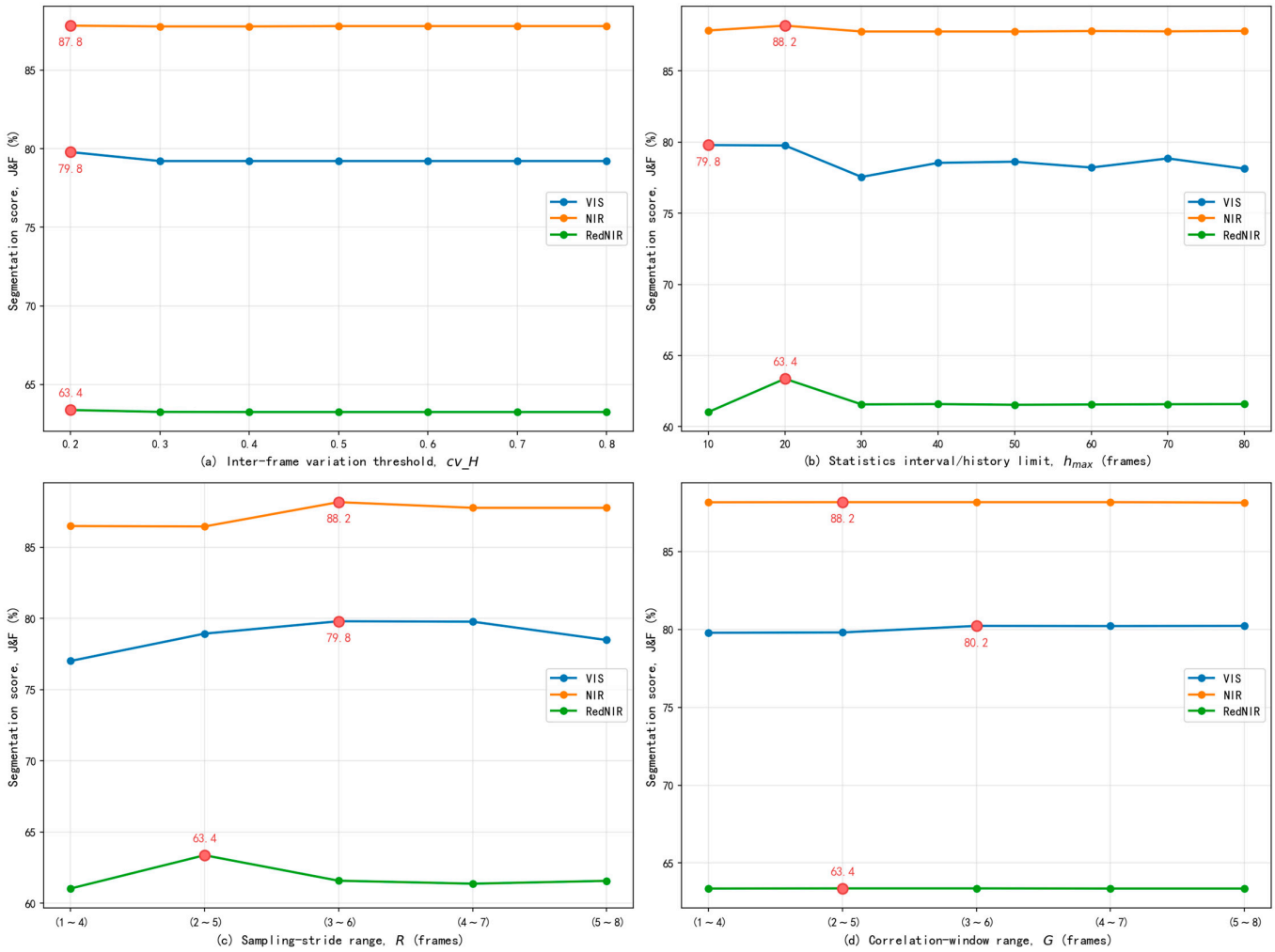


Figure 8. Parameter sensitivity analysis of ACMS on the HVOS-VIS, HVOS-NIR, and HVOS-RedNIR subsets: (**a**) inter-frame variation threshold *cv_H*; (**b**) statistics interval and maximum history length $h_{\max}$; (**c**) candidate sampling-stride range defined by *R*; and (**d**) candidate correlation-window range defined by *G*. The horizontal axes show the candidate settings of the corresponding variables, while the vertical axes report the $\mathcal{J}\&\mathcal{F}(\%)$. For *R* and *G*, a label such as "1~4" denotes the minimum-to-maximum output range of the corresponding piecewise function across different target states. The highlighted circles and adjacent values indicate the subset-specific selected settings and their corresponding scores.

As shown in Figure 8, ACMS is relatively insensitive to *cv_H* and *G*, with broadly consistent performance trends across the three subsets. A value of 0.2 is selected for *cv_H* on all three subsets. For *G*, $(2 \sim 5)$ is selected on HVOS-NIR and HVOS-RedNIR, whereas $(3 \sim 6)$ is used on HVOS-VIS. Model performance is more sensitive to $h_{\max}$ and *R*. For $h_{\max}$, the best-performing value is 20 on HVOS-NIR and HVOS-RedNIR and 10 on HVOS-VIS. For *R*, $(3 \sim 6)$ is selected on HVOS-VIS and HVOS-NIR, while $(2 \sim 5)$ is adopted on HVOS-RedNIR. These results indicate that appropriate settings of the key ACMS parameters can further improve segmentation performance across different spectral scenarios.

4.5. Discussion

The use of SAM 2 provides SSAM-SAM2 with generic spatial and temporal representation priors learned from large-scale image and video data. These pretrained representations facilitate object-level feature extraction, boundary localization, and temporal target association, thereby providing a transferable foundation for hyperspectral video object

segmentation. However, because SAM 2 is originally designed for three-channel visual inputs, its pretrained representations do not directly model the high-dimensional spectral responses of hyperspectral imagery. In the proposed framework, SSIE therefore serves as a modality-adaptation interface between hyperspectral observations and the input space of SAM 2. It selects informative spectral bands, suppresses redundant or noisy spectral responses, preserves complementary information from the residual bands, and integrates these spectral cues with the spatial details of the corresponding false-color RGB images. Nevertheless, this modality-adaptation process should not be interpreted as exhaustive modeling of the intrinsic characteristics of hyperspectral data. Once the full-band information has been encoded into the three-channel spectral–spatial representation, the downstream SAM 2 backbone no longer explicitly preserves the spectral axis or models wavelength-wise dependencies. Therefore, the improvements obtained by SSIE demonstrate the effective use of complementary spectral information within the current adaptation framework, rather than the complete exploitation of all hyperspectral characteristics. In addition, although the downstream segmentation loss provides optimization feedback to SSIE during end-to-end training, the current model performs a single feed-forward spectral adaptation during inference. It does not explicitly re-evaluate or reselect the core bands according to the downstream segmentation confidence within each forward pass. The residual-compensation and direct false-color RGB branches mitigate the influence of incomplete core-band selection, but they cannot guarantee that the selected bands are optimal under all imaging conditions.

The consistent performance improvements obtained on the VIS, NIR, and RedNIR subsets indicate that the combination of the pretrained SAM 2 representation and SSIE can accommodate, to a certain extent, differences in spectral range, band number, image appearance, and noise distribution within HVOS. These results suggest a degree of cross-spectral adaptability across the three evaluated hyperspectral settings. Nevertheless, such cross-spectral consistency should not be interpreted as conclusive evidence of cross-domain generalization to other remote sensing datasets. The three subsets are derived from the same HOT benchmark and share related data-construction and evaluation protocols. The current experiments therefore do not fully account for domain shifts caused by independent datasets, different sensor systems, acquisition platforms, spatial resolutions, environmental conditions, or spectral configurations.

The small size of the RedNIR test set, which contains only 11 video sequences, is another limitation. The selected sequences include challenging conditions such as pronounced background interference, target-scale variation, occlusion, and difficult spectral responses, making them useful for examining model behavior under demanding RedNIR scenarios. However, the limited number of sequences may increase the sensitivity of the quantitative results to the composition of the test set and restrict the statistical representativeness of the conclusions. Therefore, the results on HVOS-RedNIR should be regarded as supporting evidence within the current benchmark rather than as a definitive demonstration of broad generalization.

Another practical consideration is the model complexity. As reported in Tables 1–4, SSAM-SAM2 contains 80.86 M parameters, which is relatively large compared with some lightweight video object segmentation models. However, most of these parameters are inherited from the SAM2-B+ architecture rather than introduced by the proposed modules. The original SAM2-B+ baseline contains 80.85 M parameters, while SSIE increases the parameter count by only approximately 0.01 M through its lightweight band-scoring, band-embedding, spectral-attention, and branch-weighting operations. ACMS is training-free and introduces no additional learnable parameters. Most of the model parameters therefore originate from the Hiera-B+ image encoder and the memory-attention and mask-decoding components of SAM 2, rather than from SSIE or ACMS. In terms of computational complexity,

SSIE increases the cost from 91.8 GFLOPs for the SAM2-B+ baseline to 92.2–92.7 GFLOPs across the three subsets. This small variation is associated with the different numbers of valid spectral bands processed in VIS, NIR, and RedNIR. Although the proposed modules introduce only limited additional parameters and computations, the overall size of the SAM 2 backbone remains a constraint for deployment on resource-limited platforms. Moreover, despite introducing no learnable parameters, the online feature-correlation and dynamic memory-selection operations of ACMS may incur additional runtime and memory-access overhead that is not fully reflected by the parameter count.

Future work will evaluate the proposed method on independent hyperspectral video datasets acquired using different sensors, platforms, and environmental conditions. These experiments will examine whether the pretrained spatial priors of SAM 2 and the modality adaptation achieved by SSIE can generalize beyond the current HVOS benchmark. In parallel, lightweight SAM 2 backbones, model pruning, knowledge distillation, and more efficient memory-selection strategies will be investigated to reduce the computational and memory requirements while retaining the benefits of pretrained spatial representations and adaptive temporal modeling. To strengthen the interaction between spectral modeling and downstream segmentation, future work will also explore multi-level spectral adapters or spectral prompts inserted into intermediate SAM 2 layers, confidence-aware feedback for band re-evaluation, and HSI-native video segmentation backbones capable of directly processing full-band hyperspectral inputs.

5. Conclusions

Hyperspectral videos provide material-related spectral information that is unavailable in conventional visible-light videos, offering additional discriminative cues for separating targets from backgrounds with similar spatial appearances. However, applying existing RGB-based video object segmentation methods to hyperspectral videos remains challenging because of high-dimensional and redundant spectral information, fixed temporal memory strategies, and the limited availability of task-specific pixel-level annotations. To address these issues, this study investigated hyperspectral video object segmentation from the complementary perspectives of dataset construction, spectral–spatial representation, and adaptive temporal memory modeling.

We constructed HVOS, a hyperspectral video object segmentation dataset containing hyperspectral videos, their corresponding HSI-derived false-color RGB videos, and frame-by-frame pixel-level annotations. According to their spectral coverage and imaging characteristics, the video sequences are organized into three subsets: VIS, NIR, and RedNIR. HVOS provides a unified basis for training and evaluating video object segmentation methods under different hyperspectral settings. Based on this dataset, we developed SSAM-SAM2 by introducing SSIE and ACMS into the SAM 2 framework. SSIE selects informative spectral bands, suppresses redundant and noisy responses, preserves complementary information from the residual bands, and integrates these spectral cues with the spatial structure of the corresponding false-color RGB images. ACMS characterizes variations in target scale and temporal state and accordingly adjusts the memory-sampling stride and correlation-selection window, reducing the influence of redundant or unreliable historical information.

Comparative experiments on the three HVOS subsets show that SSAM-SAM2 outperforms the selected comparison methods under the evaluated settings. The ablation results further demonstrate that SSIE improves target representation by exploiting spectral–spatial complementarity, whereas ACMS enhances temporal stability through target-state-aware memory selection. Their combination produces consistent performance gains across VIS, NIR, and RedNIR, supporting the effectiveness of combining the pretrained representation

of SAM 2 with hyperspectral feature adaptation and target-state-aware memory management within the current HVOS benchmark.

Nevertheless, the present findings remain confined to the current HVOS benchmark, particularly the relatively small RedNIR test set. Future work will validate the proposed framework on independent hyperspectral video datasets and investigate more efficient memory-selection mechanisms and HSI-native architectures to improve computational efficiency and full-band spectral modeling.

Author Contributions: Conceptualization, B.S. and W.S.; methodology, Y.C. and Y.T.; software, Y.C.; validation, Y.C. and B.S.; formal analysis, Y.C.; investigation, B.S.; resources, W.S.; data curation, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, B.S. and W.S.; visualization, Y.C.; supervision, B.S.; project administration, B.S.; funding acquisition, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62375211.

Data Availability Statement: The Hyperspectral Object Tracking (HOT) dataset used to construct the HVOS dataset is publicly available at https://www.hsitracking.com (accessed on 14 August 2026). The pixel-level annotations developed for HVOS are available from the corresponding author upon reasonable request. The source code for SSAM-SAM2 will be made publicly available at https://github.com/ChallengeYF/SSAM-SAM2.

Acknowledgments: In the preparation of this manuscript, the authors used ChatGPT 5.6 solely for language editing and text polishing. The purpose of using this tool was to improve the grammar, clarity, and overall readability of the manuscript. No generative AI tool was used for the creation of scientific content, data analysis, data interpretation, or formulation of conclusions. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; Van Gool, L. One-shot video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5320–5329.
2. Li, H.; Zhao, X.; Su, A.; Zhang, H.; Liu, J.; Gu, G. Color space transformation and multi-class weighted loss for adhesive white blood cell segmentation. *IEEE Access* **2020**, *8*, 24808–24818. [CrossRef]
3. Zhou, S.; Nie, D.; Adeli, E.; Yin, J.; Lian, J.; Shen, D. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 461–475. [CrossRef] [PubMed]
4. Pantazi, X.E.; Moshou, D.; Tamouridou, A.A. Automated leaf disease detection in different crop species through image features analysis and one-class classifiers. *Comput. Electron. Agric.* **2019**, *156*, 96–104. [CrossRef]
5. Lee, S.Y.; Tama, B.A.; Moon, S.J.; Lee, S. Steel surface defect diagnostics using deep convolutional neural network and class activation map. *Appl. Sci.* **2019**, *9*, 5449. [CrossRef]
6. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
7. Wang, Y.; Qi, Q.; Jiang, L.; Liu, Y. Hybrid remote sensing image segmentation considering intrasegment homogeneity and intersegment heterogeneity. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 22–26. [CrossRef]
8. Zhang, L.; Ma, J.; Lv, X.; Chen, D. Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 117–121. [CrossRef]
9. Ngan, K.N.; Li, H. (Eds.) *Video Segmentation and Its Applications*; Springer: New York, NY, USA, 2011.
10. Zhang, J.; Tang, H. SAM2 for image and video segmentation: A comprehensive survey. *arXiv* **2025**, arXiv:2503.12781.
11. Gao, M.; Zheng, F.; Yu, J.J.Q.; Shan, C.; Ding, G.; Han, J. Deep learning for video object segmentation: A review. *Artif. Intell. Rev.* **2023**, *56*, 457–531. [CrossRef]
12. Vivone, G. Multispectral and hyperspectral image fusion in remote sensing: A survey. *Inf. Fusion* **2023**, *89*, 405–417. [CrossRef]

13. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. [CrossRef]
14. Zhao, D.; You, M.; Xiang, P.; Hu, J.; Asano, Y.; Yu, X.; Hsu, C.-C.; Zhou, H.; Ren, J. TBCNet: Twin-branch collaborative network for hyperspectral anomaly detection. *Pattern Recognit.* **2026**, *180*, 114647. [CrossRef]
15. Wang, Y.; Li, X.; Yang, X.; Ge, F.; Wei, B.; Li, L.; Yue, S. Deep feature-based hyperspectral object tracking: An experimental survey and outlook. *Remote Sens.* **2025**, *17*, 645. [CrossRef]
16. Qian, K.; Shen, J.; Wang, S.; Sun, W. Recent advances in object tracking using hyperspectral videos: A survey. *Multimed. Tools Appl.* **2024**, *83*, 56155–56181. [CrossRef]
17. Zhao, D.; Yu, J.; Wei, Y.; Huang, K.; Xiang, P.; Zhou, H.; Asano, Y.; Li, Y.; Arun, P.V. Depth–area estimation-based hyperspectral video tracker for scale variation adaptation. *Eng. Appl. Artif. Intell.* **2026**, *181*, 115778. [CrossRef]
18. Li, J.; Shah, I.A.; Geever, D.; Ward, E.; Glavin, M.; Jones, E.; Deegan, B. Hyperspectral vs. RGB for pedestrian segmentation in urban driving scenes: A comparative study. In Proceedings of the IEEE International Conference on Vehicular Electronics and Safety (ICVES), Coventry, UK, 27–29 October 2025; pp. 387–392.
19. 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* **2024**, arXiv:2408.00714.
20. Wu, L.; You, M.; Zhao, D.; Xu, X.; Yao, Y.; Arun, P.V.; Cheng, K.; Xiang, P. DSFNet: Dual-stream spatial–spectral fusion network for hyperspectral anomaly detection. *Digit. Signal Process.* **2026**, *175*, 106019. [CrossRef]
21. Cheng, H.K.; Schwing, A.G. XMem: Long-term video object segmentation with an Atkinson–Shiffrin memory model. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 640–658.
22. Videnović, J.; Lukežič, A.; Kristan, M. A distractor-aware memory for visual object tracking with SAM2. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 24255–24264.
23. Pourganjalikhan, A.; Poullis, C. Adaptive memory management for video object segmentation. In Proceedings of the 19th Conference on Robots and Vision (CRV), Toronto, ON, Canada, 31 May–2 June 2022; pp. 75–82.
24. Ji, W.; Li, J.; Li, W.; Shen, Y.; Cheng, L.; Jin, H. Unleashing multispectral video's potential in semantic segmentation: A semi-supervised viewpoint and new UAV-view benchmark. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 65717–65737. [CrossRef]
25. Kang, J.; Zhang, Y.; Liu, X.; Cheng, Z. Hyperspectral image classification using spectral–spatial double-branch attention mechanism. *Remote Sens.* **2024**, *16*, 193. [CrossRef]
26. Arshad, T.; Zhang, J.; Anyembe, S.C.; Mehmood, A. Spectral spatial neighborhood attention transformer for hyperspectral image classification. *Can. J. Remote Sens.* **2024**, *50*, 2347631. [CrossRef]
27. Lu, Y.; Zhang, Y.; Jiang, X.; Liu, X.; Cai, Z. Dual-branch convolution–transformer network with spectral–spatial attention for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5518216. [CrossRef]
28. Wang, W.; Sun, Q.; Zhang, L.; Ren, P.; Wang, J.; Ren, G.; Liu, B. A spatial–spectral fusion convolutional transformer network with contextual multi-head self-attention for hyperspectral image classification. *Neural Netw.* **2025**, *187*, 107350. [CrossRef] [PubMed]
29. Chen, Y.; Yuan, Q.; Xie, H.; Tang, Y.; Xiao, Y.; He, J.; Guan, R.; Liu, X.; Zhang, L. Hyperspectral video tracking with spectral–spatial fusion and memory enhancement. *IEEE Trans. Image Process.* **2025**, *34*, 3547–3562. [CrossRef] [PubMed]
30. Zhao, D.; Zhang, H.; Huang, K.; Zhu, X.; Arun, P.V.; Jiang, W.; Li, S.; Pei, X.; Zhou, H. SASU-Net: Hyperspectral video tracker based on spectral adaptive aggregation weighting and scale updating. *Expert Syst. Appl.* **2025**, *272*, 126721. [CrossRef]
31. Chen, Y.; Yuan, Q.; Tang, Y.; Xiao, Y.; He, J.; Han, T.; Liu, Z.; Zhang, L. SSTtrack: A unified hyperspectral video tracking framework via modeling spectral–spatial–temporal conditions. *Inf. Fusion* **2025**, *114*, 102658. [CrossRef]
32. Liu, S.; Guo, H.; Gao, S.; Zhang, W. The spectrum difference enhanced network for hyperspectral anomaly detection. *Remote Sens.* **2024**, *16*, 4518. [CrossRef]
33. Yao, Y.; Wang, Q.; Zhao, D.; You, M.; Xiang, P.; Asano, Y.; Yu, X.; Wang, C.; Zhou, H.; Ren, J. DFBSNet: Dual frequency-domain branch fusion and selection network for hyperspectral anomaly detection. *Pattern Recognit.* **2026**, *180*, 113967. [CrossRef]
34. Shao, Y.; Li, S.; Yang, P.; Cheng, F.; Ding, Y.; Sun, J. JointNet: Multitask learning framework for denoising and detecting anomalies in hyperspectral remote sensing. *Remote Sens.* **2024**, *16*, 2619. [CrossRef]
35. You, M.; Yao, Y.; Zhao, D.; Zhao, Z.; Arun, P.V.; Wang, Y.; Zhou, H.; Chi, R. S3CRAD: Superpixel-guided background inpainting and spatial–spectral constrained representation for hyperspectral anomaly detection. *Opt. Lasers Eng.* **2026**, *201*, 109657. [CrossRef]
36. Voigtlaender, P.; Leibe, B. Online adaptation of convolutional neural networks for video object segmentation. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017.
37. Ding, H.; Ying, K.; Liu, C.; He, S.; Jiang, X.; Jiang, Y.-G.; Torr, P.H.S.; Bai, S. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv* **2025**, arXiv:2508.05630.
38. Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; Sorkine-Hornung, A. Learning video object segmentation from static images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3491–3500.

39. Oh, S.W.; Lee, J.-Y.; Sunkavalli, K.; Kim, S.J. Fast video object segmentation by reference-guided mask propagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7376–7385.
40. Yang, Z.; Wei, Y.; Yang, Y. Associating objects with transformers for video object segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2491–2502.
41. Yang, Z.; Yang, Y. Decoupling features in hierarchical propagation for video object segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36324–36336. [CrossRef]
42. 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 (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9225–9234.
43. Cheng, H.K.; Tai, Y.-W.; Tang, C.-K. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 11781–11794.
44. Bekuzarov, M.; Bermudez, A.; Lee, J.-Y.; Li, H. XMem++: Production-level video segmentation from few annotated frames. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 635–644.
45. Cheng, H.K.; Oh, S.W.; Price, B.; Lee, J.-Y.; Schwing, A.G. Putting the object back into video object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 3151–3161.
46. Li, M.; Hu, L.; Xiong, Z.; Zhang, B.; Pan, P.; Liu, D. Recurrent dynamic embedding for video object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1322–1331.
47. Zhang, J.; Cui, Y.; Wu, G.; Wang, L. JointFormer: A unified framework with joint modeling for video object segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 6039–6054. [CrossRef] [PubMed]
48. Mansoori, M.; Shahabodini, S.; Abouei, J.; Plataniotis, K.N.; Mohammadi, A. Self-prompting polyp segmentation in colonoscopy using hybrid YOLO-SAM2 model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
49. Liu, H.; Zhang, E.; Wu, J.; Hong, M.; Jin, Y. Surgical SAM 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv* **2024**, arXiv:2408.07931.
50. Wang, Z.; Zhou, Q.; Liu, Z. Det-SAM2: Technical report on the self-prompting segmentation framework based on Segment Anything Model 2. *arXiv* **2024**, arXiv:2411.18977.
51. 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. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 13614–13624.
52. Qian, Z.; Li, Z.; Feng, B.; Liu, W.; Wang, X. Track hyperspectral videos via Segment Anything Model 2. In Proceedings of the 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Helsinki, Finland, 9–11 December 2024; pp. 1–5.
53. He, W.; Ren, Q.; Zhang, H.; Mo, Z.; Li, N.; Xu, M.; Jia, S. DeSAM2: Detection-driven SAM2 for hyperspectral object tracking. In Proceedings of the 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Helsinki, Finland, 9–11 December 2024; pp. 1–5.
54. 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 Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 29441–29454.
55. Tan, Y.; Sun, W.; Li, J.; Hou, S.; Li, X.; Wang, Z.; Song, B. HyperTrack: A unified network for hyperspectral video object tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 1015–1028. [CrossRef]
56. Gao, L.; Zhang, Y.; Chen, L.; Jiang, Y.; Xie, W.; Li, Y. Hyperspectral adapter for object tracking based on hyperspectral video. *arXiv* **2025**, arXiv:2503.22199.
57. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.
58. Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; Huang, T. YouTube-VOS: Sequence-to-sequence video object segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 585–601.
59. Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P.H.S.; Bai, S. MOSE: A new dataset for video object segmentation in complex scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 20167–20177.

60. Li, Z.; Xiong, F.; Lu, J.; Wang, J.; Chen, D.; Zhou, J.; Qian, Y. Multi-domain universal representation learning for hyperspectral object tracking. *Pattern Recognit*. **2025**, *162*, 111389.
61. Xiong, F.; Wu, Z.; Zhou, J.; Jia, S.; Qian, Y. SUIT: Spatial-spectral union-intersection interaction network for hyperspectral object tracking. *IEEE Trans. Image Process*. **2025**, *34*, 7786–7800.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.