Hyper-graph Video Object Segmentation via Text-depth Collaborative Reasoning

Jiaqing Fan, Yifan Liao, Fanzhang Li*
School of Computer Science & Technology, Soochow University
jqfan@suda.edu.cn, 20234227048@stu.suda.edu.cn, lfzh@suda.edu.cn

Abstract

Current video object segmentation (VOS) solutions often overlook the wealthy information *e.g.* subtitles and depth cues among video sequences, which are crucial for effectively linking video content. Recognizing the significance of these elements, in this paper, we introduce a novel approach termed as "Hyper-graph Text-Depth Collaborative Reasoning Video Object Segmentation" (HTD). It aims to leverage the synergy between textual and depth information to enhance the segmentation of objects in video sequences. The HTD framework integrates textual and depth data into a hyper-graph structure, where nodes represent objects, text, and depth features, and hyper-edges encode complex relationships among them. After grabbing the multimodal context of video scenes, the proposed collaborative reasoning mechanism within the hyper-graph iteratively refines object boundaries by considering the interplay between textual cues, depth information, and spatial-temporal coherence. We demonstrate the effectiveness of HTD through extensive experiments on four benchmarks. The results show that our approach outperforms state-of-the-art VOS methods, particularly in scenarios with complex backgrounds, occlusions, and dynamic scenes. The inclusion of text and depth data not only improves segmentation accuracy but also enhances the interpretability of the segmentation process. We have released the training and testing code on https://github.com/zyaireleo/HTD.git.

1 Introduction

Video Object Segmentation (VOS) is a fundamental research problem in computer vision, aiming to accurately identify and segment specific target objects in video sequences. With the rapid development of deep learning technologies, VOS has shown broad application prospects across various fields: in video editing and post-production, it enables automatic matting and special effects addition; in autonomous driving, VOS helps vehicles accurately identify and track various obstacles on the road; in video surveillance systems, it enables continuous tracking of specific targets; in augmented reality applications, VOS technology supports precise integration of real objects with virtual content; in medical image analysis, it can assist doctors in tracking and analyzing specific tissues or organs in dynamic medical images [Xu *et al.*, 2023; Peng *et al.*, 2025; Wu *et al.*, 2024]. As a key technology, VOS is continually advancing our capabilities in visual content processing and understanding. However, due to the complexity of video scenes, uncertainty of object motion, and various occlusion and interference factors, achieving accurate and stable video object segmentation still faces challenges.

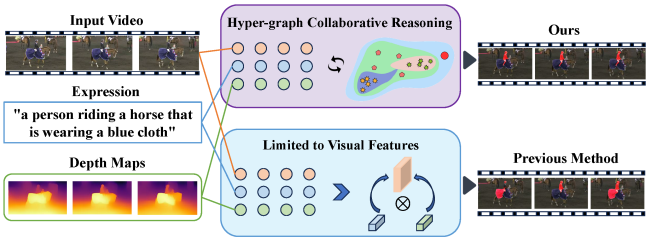

Figure 1: By constructing a multimodal hyper-graph structure, our solution enables collaborative reasoning among textual expressions, depth maps, and visual features for VOS. Compared to traditional methods that heavily rely on visual features, it achieves better results, particularly excelling in complex background scenarios.

Existing VOS methods [Liang *et al.*, 2023; Wu *et al.*, 2022; Luo *et al.*, 2023; Miao *et al.*, 2023a; Yang *et al.*, 2024; He and Ding, 2024] primarily rely on visual features and spatio-temporal consistency information from RGB image sequences, but they often overlooking other rich modal information contained in videos, such as subtitle text and depth information. This additional modal information holds significant value for understanding video content and accurately segmenting target objects. Particularly in challenging situations involving complex backgrounds, object occlusions, and dynamic scenes, relying solely on visual features often fails to achieve satisfactory segmentation results. For instance, when target objects have similar appearance features to the background, introducing depth information can help better distinguish foreground from background; meanwhile, semantic information contained in subtitle text can provide important clues for object identity recognition and cross-shot tracking.

*Corresponding author.

Based on these observations, as illustrated in figure 1, we propose a novel Hyper-graph Text-Depth Collaborative Reasoning Video Object Segmentation (HTD) method. The core idea of this method is to construct a multimodal hyper-graph structure that uniformly represents target objects, textual information, and depth features in videos as graph nodes, while using hyper-edges to describe their complex relationships. Through iterative reasoning on the hyper-graph, our method can fully utilize the interactions between textual cues, depth information, and spatio-temporal consistency to continuously optimize object boundary segmentation results. Compared to existing methods, our proposed HTD framework offers the following advantages: first, it can effectively integrate multimodal information, fully exploiting textual and depth cues in video sequences. Second, the hyper-graph-based collaborative reasoning mechanism enables deep fusion and interaction of various modal information, thereby producing more accurate segmentation results. Finally, the introduction of text and depth information not only improves segmentation accuracy but also enhances the interpretability of the segmentation process. Extensive experiments demonstrate that our method achieves superior performance to existing state-of-the-art methods on four mainstream datasets, showing particularly significant advantages in handling challenging scenarios. In conclusion, the presented method represents a significant advancement in video object segmentation, providing a robust framework that can effectively utilize multimodal information to achieve superior segmentation performance. The contributions are summarized as follows:

- We introduce a novel approach termed as "Hyper-graph Text-Depth Collaborative Reasoning Video Object Segmentation" (HTD), which integrates appearance, location, text, and depth data to enhance VOS, offering more context for accurate video object segmentation.
- We propose a progressive hyper-graph model that integrated modalities such as appearance, location, text, and depth to construct hyper-edges between spatially and temporally adjacent regions. It also incorporates multiple object proposals per frame, rather than relying on a single object proposal per frame, thereby generating more reliable object regions per frame.
- Collaborative reasoning mechanism that takes into account higher-order correlations, effectively handles complex non-pairwise relationships in video frames, and simultaneously integrates them together.
- HDT yields favorable performance on four challenging benchmarks. The inclusion of text and depth data not only improves segmentation accuracy but also enhances the interpretability of the segmentation process.

2 Related Works

2.1 Video Object Segmentation

Video object segmentation is a research hotspot in computer vision, and great progress has been made in semi-supervised, unsupervised and referring VOS fields in recent years. **Semi-supervised VOS** aims to utilize the segmentation annotation of the first frame to segment the target in subsequent video frames. In recent years, researchers have proposed a series of innovative methods to enhance the performance of semi-supervised VOS: Cheng et.al. [Cheng *et al.*, 2024] propose Cutie that introduces object-level memory reading through object queries, contrasting with traditional pixel-level approaches. It uses a query-based object transformer and foreground-background masked attention to better separate target objects from backgrounds. On the popular datasets, Cutie achieves significant improvements, while running 3x faster. Li et.al. [Li *et al.*, 2025] introduce a novel unified framework for Video Object Segmentation that integrates feature extraction, matching, memory management, and multi-object aggregation into a single transformer-based architecture. Unlike traditional approaches that handle these components separately, OneVOS models all features as transformer tokens and processes them through a unified attention mechanism. **Unsupervised VOS** does not require any manual labeling, but automatically finds and segments the main moving objects in the video. Traditional methods mainly rely on low-level clues such as optical flow and moving boundary detection. In recent years, deep learning methods have made important breakthroughs in this field. Cho et.al. [Cho *et al.*, 2024] introduce two novel attention mechanisms (inter-modality attention and inter-frame attention) to improve unsupervised video object segmentation by better integrating appearance and motion information across frames and modalities. Ding et.al. [Ding *et al.*, 2025] introduce a simplified self-supervised video object segmentation approach that leverages DINO-pretrained Transformers' inherent objectness and spatio-temporal dependencies, achieving state-of-the-art results without requiring auxiliary modalities or iterative slot attention. **Referring VOS** are designed to locate and segment specific targets in a video based on linguistic references. This is a multimodal understanding problem: Botach et.al. [Botach *et al.*, 2022a] present MTTR (Multimodal Tracking Transformer), a simple yet effective end-to-end Transformer-based approach that treats referring video object segmentation as a sequence prediction problem, achieving state-of-the-art performance with a significantly simplified pipeline. Miao et.al. [Miao *et al.*, 2023b] propose a Spectrum-guided Multi-granularity (SgMg) approach that addresses feature drift in referring video object segmentation by performing direct segmentation on encoded features and introducing spectral domain fusion, enabling efficient multi-object segmentation with state-of-the-art performance. Zhu et.al. [Zhu *et al.*, 2025] introduce VD-IT, a framework that leverages pre-trained text-to-video diffusion models for referring video object segmentation, demonstrating that generative models can maintain better alignment and temporal consistency compared to traditional discriminative backbones.

2.2 Hyper-graph Learning

The development of hyper-graph learning has gone through several important stages: Beginning in the early 2000s with spectral theory-based research, researchers extended traditional graph spectral theory to the hyper-graph domain, proposing hyper-graph Laplacian matrices and spectral clustering algorithms, laying the theoretical foundation for hyper-graph data analysis [Chan *et al.*, 2018]. Around 2010,

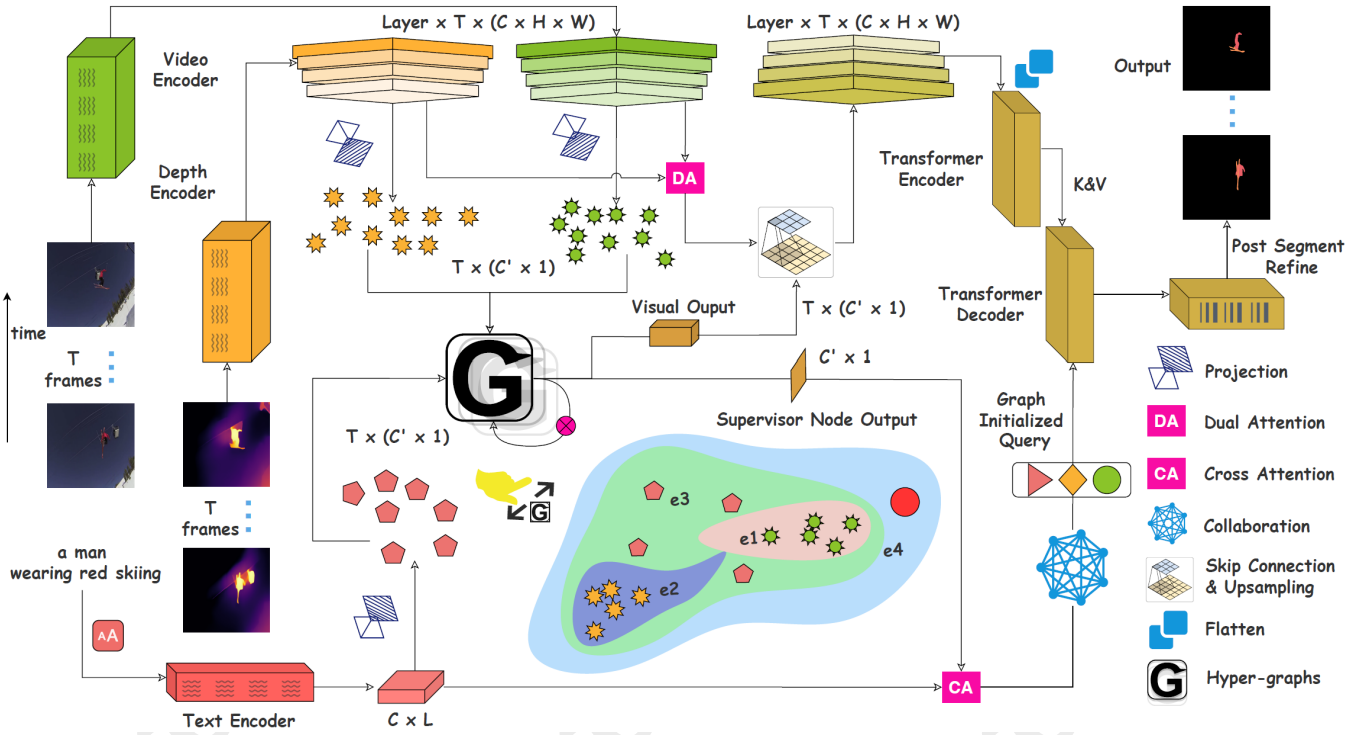

Figure 2: Overall architecture of the proposed HTD method: The approach extracts multimodal features through video encoder, depth encoder, and text encoder respectively, constructing a hyper-graph structure that incorporates target objects, textual information, and depth features. Through dual attention and cross attention mechanisms, it enables collaborative reasoning of multimodal features, utilizes Transformer encoder-decoder for feature enhancement, and achieves precise object segmentation results after post-processing optimization. This architecture effectively synthesizes multimodal information, demonstrating significant advantages particularly in complex backgrounds.

research focus shifted towards tensor decomposition-based methods, capturing complex interactions between multiple entities through tensor decomposition techniques, while also developing random walk-based hyper-graph representation learning methods [Antelmi *et al.*, 2023]. After 2015, the rise of deep learning brought revolutionary changes, with researchers extending graph neural networks to the hyper-graph domain [Kim *et al.*, 2020], proposing innovative architectures such as hyper-graph neural networks, and developing models like hyper-graph attention networks and dynamic hyper-graph neural networks. Research after 2020 has increasingly focused on the dynamics and heterogeneity of hyper-graphs [Fan *et al.*, 2021], beginning to explore self-supervised learning in the hyper-graph domain, while emphasizing the interpretability and robustness of hyper-graphs.

3 Methodology

As illustrated in figure 2, we propose HTD (Hyper-graph Transformer Dual-stream), a novel framework for video object segmentation that leverages multi-modal information through hyper-graph reasoning and Transformer-based feature synthesis. Our approach consists of four main components: multi-modal text-depth video encoding, hyper-graph construction and reasoning, Transformer collaborative synthesis, and post-processing refinement. Figure 2 illustrates the overall architecture of our proposed method.

3.1 Multi-modal Text-Depth Video Encoding

Video Encoder. The video encoder processes a sequence of T frames to extract rich spatio-temporal features. For each frame, we employ a hierarchical CNN architecture that generates feature maps with dimensions Layer × T × (C × H × W), where Layer denotes the number of hierarchical features, T represents the temporal dimension, and C, H, W correspond to the channel, height, and width dimensions respectively. This multi-scale representation captures both fine-grained details and high-level semantic information across different temporal stages.

Depth Encoder. To incorporate geometric cues, we introduce a dedicated depth encoder that processes the corresponding depth information for each frame. The depth encoder maintains the same architectural design as the video encoder, producing feature maps with matching dimensions. This parallel processing ensures geometric information can be effectively integrated with appearance features in subsequent stages.

Text Encoder. Natural language descriptions provide crucial semantic guidance for identifying target objects. Our text encoder converts the input text query (e.g., "a man wearing red skiing") into a dense feature representation of dimension C × L, where L denotes the sequence length. We utilize a pre-trained transformer-based language model fine-tuned on our task to generate contextually rich text embeddings.

3.2 Hyper-graph Construction and Reasoning

Node Construction. The foundation of our multi-modal reasoning framework is a dynamic hyper-graph structure that serves as a unified representation space where features from different modalities can effectively interact. In the node construction phase, we project visual (F_v), depth (F_d), and textual (F_t) features into a common embedding space through learnable linear transformations:

$$N_i = W_i F_i + b_i, \quad i \in \{v, d, t\}, \tag{1}$$

where W_i and b_i are modality-specific projection parameters. Each node represents a semantic unit that can participate in cross-modal reasoning.

Hyper-edge Formation. The hyper-graph structure is defined through four distinct types of hyper-edges, each designed to capture specific aspects of multi-modal relationships. Spatio-temporal relations (e_1) connect nodes across spatial and temporal dimensions, enabling the model to reason about object consistency and motion patterns. The connection strength α_{st} between nodes i and j is computed as:

$$\alpha_{st}(i,j) = \text{Softmax}(\frac{(W_q N_i)^T (W_k N_j)}{\sqrt{d}} + \beta_{ij}), \tag{2}$$

where β_{ij} represents the spatial proximity bias and d is the feature dimension.

Text-visual associations (e_2) bridge the semantic gap between language and vision through hyper-edges that link text feature nodes with relevant visual feature nodes. The association strength is determined using a cross-modal attention mechanism that measures semantic compatibility. Depth-region correspondences (e_3) establish connections between depth feature nodes and spatial regions in the visual features, incorporating geometric constraints for improved segmentation accuracy. Finally, global context hyper-edges (e_4) connect all nodes within each modality, facilitating long-range dependency modeling and holistic scene understanding.

Iterative Reasoning. The hyper-graph reasoning process follows an iterative message passing scheme where node features are progressively refined through multiple rounds of information exchange. During message generation, for each hyper-edge type e_k, we aggregate information from connected nodes using modality-specific attention mechanisms:

$$M_i^k = \sum_{j \in \mathcal{N}_k(i)} \alpha_{ij}^k W_v^k N_j, \tag{3}$$

where $\mathcal{N}_k(i)$ denotes the neighboring nodes of node i connected by hyper-edge type k. The node features are then updated through a combination of self-attention and feed-forward networks:

$$N_i^{t+1} = \text{FFN}(\text{LayerNorm}(N_i^t + \sum_k \gamma_k M_i^k)), \tag{4}$$

where γ_k are learnable weights that balance the contribution of different hyper-edge types, and t denotes the iteration step. This iterative reasoning process is performed multiple times to achieve deep cross-modal fusion, enabling the model to progressively refine its understanding of the multi-modal relationships and generate more accurate segmentation results.

3.3 Transformer Collaborative Synthesis

Our transformer-based collaborative synthesis framework integrates multiple attention mechanisms to effectively fuse and enhance multi-modal features. In the feature enhancement stage, we implement two complementary attention mechanisms. The Dual Attention (DA) module operates independently on each modality to capture both spatial and channel-wise dependencies. For a given feature map $F \in \mathbb{R}^{C \times H \times W}$, the dual attention output is computed as:

$$F_{DA} = \alpha_s \cdot \text{SpatialAtt}(F) + \alpha_c \cdot \text{ChannelAtt}(F), \tag{5}$$

where α_s and α_c are learnable parameters that balance the contribution of each attention component. The Cross Attention (CA) mechanism facilitates direct interaction between different modality features, enabling effective alignment and fusion of complementary information. For features from two modalities F_1 and F_2, the cross attention is formulated as:

$$\text{CA}(F_1, F_2) = \text{Softmax}(\frac{Q_1 K_2^T}{\sqrt{d_k}}) V_2, \tag{6}$$

where Q_1, K_2, and V_2 represent the query, key, and value projections respectively, and d_k is the dimension of the key vectors.

The query initialization process builds upon the hyper-graph node outputs to preserve the rich multi-modal relationships established during hyper-graph reasoning. We employ a learnable projection layer ϕ to transform node features N into query representations Q_{init}: $Q_{init} = \phi(N) + \text{PE}$, where PE denotes positional encoding that injects spatial information into the queries.

Our transformer architecture adopts an encoder-decoder design optimized for multi-modal synthesis. The encoder processes the enhanced features through multiple self-attention layers, effectively capturing long-range dependencies and contextual relationships. The decoder then generates refined feature representations by attending to encoder outputs through cross-attention mechanisms. This architecture is specifically designed to maintain precise spatial information while synthesizing multi-modal cues. The final output feature F_{out} at each decoder layer l is computed as:

$$F_{out}^l = \text{FFN}(\text{LayerNorm}(\text{CA}(Q^l, K_{enc}, V_{enc}) + \text{SA}(Q^l))), \tag{7}$$

where FFN represents a feed-forward network, SA denotes self-attention, and K_{enc} and V_{enc} are the encoder's key and value projections respectively.

3.4 Post-processing Refinement

A critical aspect of our framework is the post-processing refinement stage, which focuses on enhancing the quality of initial segmentation results. The first key component is resolution enhancement, which employs a dual-stream architecture to recover fine-grained details that may have been lost during initial processing. We implement skip connections that combine features F_l from earlier layers with upsampled features F_u through an adaptive fusion mechanism, and yield:

Method	Publication	Backbone	Ref-YouTube-VOS			Ref-DAVIS17		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MTTR [Botach *et al.*, 2022b]	CVPR'22	Video-Swin-T	53.3	54.0	56.6	-	-	-
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	56.0	54.8	57.3	-	-	-
SgMg [Miao *et al.*, 2023a]	ICCV'23	Video-Swin-T	58.9	57.7	60.0	56.7	53.3	60.0
SOC [Luo *et al.*, 2023]	NeurIPS'23	Video-Swin-T	59.2	57.8	60.5	59.0	55.4	62.6
HTR [Miao *et al.*, 2024]	TCSVT'24	Video-Swin-T	59.8	58.3	61.3	57.2	53.8	60.6
HTD (Ours)	This work	Video-Swin-T	**59.8**	**58.4**	**61.1**	**58.6**	**55.6**	**61.6**
Pre-training with Refcoco								
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	59.4	58.0	60.9	59.7	56.6	62.8
TempCD [Tang *et al.*, 2023]	ICCV'23	Video-Swin-T	62.3	60.5	64.0	62.2	59.3	65.0
OnlineRefer [Wu *et al.*, 2023]	ICCV'23	Video-Swin-T	62.9	61.0	64.7	62.4	59.1	65.6
SgMg [Miao *et al.*, 2023a]	ICCV'23	Video-Swin-T	62.0	60.4	63.5	61.9	59.0	64.8
LAVT [Yang *et al.*, 2024]	TPAMI'24	Video-Swin-T	60.9	59.4	62.5	-	-	-
HTD (Ours)	This work	Video-Swin-T	**63.1**	**61.8**	**64.3**	**63.1**	**60.3**	**66.0**

Table 1: Quantitative comparison to methods on Ref-YouTube-VOS and Ref-DAVIS17.

F_{fused}. Our progressive upsampling strategy gradually restores the feature resolution while maintaining semantic consistency. This is followed by feature projection, where the refined features are projected onto the segmentation space through convolutional layers. To ensure stable training and effective feature transformation, we incorporate residual connections and normalization layers throughout this process.

Mask generation employs a binary classification head to predict pixel probabilities, controlled by a confidence-aware prediction function:

$$P(x,y) = \sigma(\phi(F_{fused}(x,y))) \cdot \exp(-\lambda \cdot D(x,y)), \quad (8)$$

where σ is the sigmoid function, ϕ is the classification network, $D(x,y)$ is the distance transform value, and λ controls the spatial distance influence. We train the model by combining Focal loss, Dice loss, and Cross-entropy loss. Focal loss addresses class imbalance:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t). \quad (9)$$

Dice loss maximizes the overlap between predicted and ground truth masks:

$$DL = 1 - \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}. \quad (10)$$

Cross-entropy measures the difference between predicted probability p and true distribution q: $CE = -\sum_i^N q_i \log(p_i)$. Post-processing includes boundary refinement and small object filtering to generate high-quality video segmentation results. In VOS tasks, Focal loss effectively handles foreground-background sample imbalance, Dice loss directly optimizes the IoU metric, and cross-entropy provides pixel-level supervision. Joint training with multiple loss functions comprehensively improves segmentation performance, particularly suitable for Dense Prediction problems in complex scenarios like VOS.

4 Experimental Results

4.1 Datasets and Metrics

Datasets and Settings. We evaluate our method on four prominent RVOS benchmarks: A2D-Sentences [Gavrilyuk *et al.*, 2018], Refer-YouTube-VOS [Seo *et al.*, 2020], JHMDB-Sentences [Jhuang *et al.*, 2013a], and Refer-DAVIS17 [Khoreva *et al.*, 2019]. A2D-Sentences contains 3,754 videos (3,017 for training, 737 for testing) with pixel-wise masks in three frames per video and 6,655 unique text expressions. Refer-YouTube-VOS includes 3,978 videos and 15,009 expressions, with instance masks for every fifth frame. JHMDB-Sentences and Refer-DAVIS17 extend JHMDB [Jhuang *et al.*, 2013b] and DAVIS17 [Pont-Tuset *et al.*, 2017], respectively, featuring 928 and 90 videos annotated with 1,544 expressions in total. Following [Wu *et al.*, 2022; Miao *et al.*, 2023a], we use the overall IoU, mean IoU, mAP and precision@K for A2D-Sentences and JHMDB-Sentences. For Refer-YouTube-VOS and Refer-DAVIS17, we adopt J, F, and their average $J\&F$ as metrics.

Implementation details. We extract depth features using [Ranftl *et al.*, 2020] and adopt two strategies inspired by [Wu *et al.*, 2022; Luo *et al.*, 2023]. For training without pre-training, we train the model on Ref-YouTube-VOS and directly evaluate it on its validation set and Ref-DAVIS17. When using pre-trained models, we first train on RefCOCO, RefCOCO+, and RefCOCOg [Mao *et al.*, 2016; Yu *et al.*, 2016], then fine-tune on Ref-YouTube-VOS, and evaluate on Ref-DAVIS17. Similarly, we pre-train on A2D-Sentences, fine-tune on JHMDB-Sentences, and evaluate accordingly. We use the Video Swin-Tiny [Liu *et al.*, 2022] backbone for vision and depth features in all experiments.

4.2 Quantitative Results

Ref-YouTube-VOS and Ref-DAVIS17. As reported in Table 1, experimental results demonstrate that the proposed HTD method achieves significant performance improvements in video object segmentation tasks. Specifically, without pre-

Method	Publications	Backbone	Precision					IoU		mAP
			P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
MTTR [Botach *et al.*, 2022b]	CVPR'22	Video-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	76.0	72.2	65.4	49.8	17.9	72.3	64.1	48.6
SOC [Luo *et al.*, 2023]	NeurIPS'23	Video-Swin-T	79.0	75.6	68.7	53.5	19.5	74.7	66.9	50.4
LAVT [Yang *et al.*, 2024]	TPAMI'24	Video-Swin-T	77.3	73.2	65.0	49.0	17.3	74.4	65.9	-
HTD (Ours)	This work	Video-Swin-T	**78.3**	**76.1**	**69.7**	**55.0**	**20.9**	**75.7**	**66.9**	**51.4**
Pre-training with Refcoco										
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
SOC [Luo *et al.*, 2023]	NeurIPS'23	Video-Swin-T	83.1	80.6	73.9	57.7	21.8	78.3	70.6	54.8
SgMg [Miao *et al.*, 2023a]	ICCV'23	Video-Swin-T	-	-	-	-	-	78.0	70.4	56.1
HTML [Han *et al.*, 2023]	ICCV'23	Video-Swin-T	82.2	79.2	72.3	55.3	20.1	77.6	69.2	53.4
LAVT [Yang *et al.*, 2024]	TPAMI'24	Video-Swin-T	82.8	79.3	71.5	54.6	19.5	77.9	70.0	-
TCE-RVOS [Hu *et al.*, 2024]	WACV'24	Video-Swin-T	83.0	79.9	73.6	56.7	20.5	77.5	69.9	54.8
HTD (Ours)	This work	Video-Swin-T	**84.0**	**81.0**	**75.9**	**60.1**	**23.1**	**78.6**	**70.9**	**57.0**

Table 2: Quantitative comparison to methods on A2D-Sentences.

Method	Publications	Backbone	Precision					IoU		mAP
			P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	93.3	84.2	61.4	16.4	0.3	70.0	69.3	39.1
LOCATER [Liang *et al.*, 2023]	TPAMI'23	Video-Swin-T	93.6	85.9	61.9	16.8	0.3	70.8	69.6	39.4
SOC [Luo *et al.*, 2023]	NeurIPS'23	Video-Swin-T	94.7	86.4	62.7	17.9	0.1	70.7	70.1	39.7
HTD (Ours)	This work	Video-Swin-T	**95.1**	**87.3**	**63.5**	**19.0**	**0.2**	**71.3**	**71.0**	**40.8**
Pre-training with Refcoco										
ReferFormer [Wu *et al.*, 2022]	CVPR'22	Video-Swin-T	95.8	89.3	66.8	18.9	0.2	71.9	71.0	42.2
SOC [Luo *et al.*, 2023]	NeurIPS'23	Video-Swin-T	96.3	88.7	67.2	19.6	0.1	72.7	71.6	42.7
SgMg [Miao *et al.*, 2023a]	ICCV'23	Video-Swin-T	-	-	-	-	-	72.8	71.7	44.4
DsHmp [He and Ding, 2024]	CVPR'24	Video-Swin-T	-	-	-	-	-	73.1	72.1	44.9
HTD (Ours)	This work	Video-Swin-T	**96.1**	**90.0**	**68.3**	**21.9**	**0.3**	**73.7**	**72.6**	**45.0**

Table 3: Quantitative comparison to methods on JHMDB-Sentences.

training, HTD achieves a $J\&F$ score of 59.8 on the Ref-YouTube-VOS dataset, outperforming existing methods. This result validates the effectiveness of the multi-modal data integration strategy, which provides richer contextual information by fusing appearance, location, text, and depth information. Notably, when Refcoco pre-training is introduced, HTD's performance shows substantial improvement, with the $J\&F$ score increasing from 59.8 to 63.1, fully demonstrating the advantages of the progressive hyper-graph model in handling multiple object proposals. Experiments on the Ref-DAVIS17 dataset further validate the effectiveness of the collaborative reasoning mechanism, with F-scores reaching 61.6 and 66.0 under non-pre-trained and pre-trained conditions respectively, demonstrating the mechanism's powerful capability in handling complex non-pairwise relationships between video frames. Although having only **85M** parameters, the presented approach HTD has attained a real-time speed of **50** FPS and surpassed six state-of-the-art methods from 2024 as shown in Tables above, demonstrating that our proposed solution achieves both high accuracy and efficiency with significant performance advantages. In addition, under the same setting of Video-Swin-Tiny backbone and similar model parameter count, the presented HTD achieves the **highest** performance on four mainstream datasets, including JHMDB-Sentences, A2D-Sentences, Refer-DAVIS17, and Refer-YouTube-VOS.

A2D-Sentences. In Table 2,Extensive experiments on the A2D-Sentences dataset demonstrate the superior performance of our proposed HTD method. Without pre-training, HTD achieves substantial improvements across all precision metrics, reaching 78.3% for P0.5, 20.9% for P0.9, and 75.7% for Overall IoU, consistently outperforming existing approaches. When incorporating Refcoco pre-training, the performance is further enhanced, with P0.5 increasing to 84.0%, P0.9 reaching 23.1%, and Overall IoU improving to 78.6%. The significant performance gain validates the effectiveness of our hyper-graph structure in encoding multi-modal feature relationships and the synergistic effect of integrating textual and depth information. Moreover, the improved mAP score of 57.0 indicates the method's robust performance across various scenarios, particularly in handling complex backgrounds and occlusions, which aligns with our

Figure 3: Qualitative examples from two statements of one video.

design objectives. These comprehensive results demonstrate that our HTD framework successfully leverages the complementary strengths of textual cues, depth information, and spatial-temporal coherence to achieve state-of-the-art performance in video object segmentation tasks.

JHMDB-Sentences. As showing in Table 3, the extensive experiments on the JHMDB-Sentences dataset further validate the effectiveness and superiority of our proposed HTD framework. The results demonstrate that even without pre-training, HTD achieves remarkable performance with 95.1% P0.5 and 71.3% Overall IoU, surpassing existing methods and confirming the advantages of our multi-modal data integration strategy that combines appearance, location, text, and depth information. When incorporating Refcoco pre-training, the model's performance is further enhanced and achieves favorable performance, with P0.5 increasing to 96.1% and Overall IoU improving to 73.7%, strongly validating the effectiveness of the presented progressive hyper-graph model in capturing spatio-temporal relationships in video sequences.

4.3 Ablation Study

In Table 4, to thoroughly understand the role of each component in the HTD framework, we conducted detailed ablation experiments on the Ref-YouTube-VOS dataset. The results show that the complete HTD model (Version VIII) achieves optimal performance ($J\&F$=59.8), which stems from the effective synergy of various innovative components. Specifically, the importance of multi-modal data integration is validated through comparisons between the complete model and single-modality versions: when removing depth information (Version II, $J\&F$=59.7) or retaining only text information (Version III, $J\&F$=59.6), model performance decreases, confirming the necessity of multi-modal fusion strategy. The contribution of the progressive hyper-graph model is reflected in the performance difference between Version I ($J\&F$=59.5) and the baseline (Version VII, $J\&F$=58.9), with a 0.6 percentage point improvement clearly demonstrating the advantages of the hyper-graph structure in modeling spatio-temporal relationships among texts, depths, and frames.

We conducted ablation experiments combining different loss functions to evaluate the impact of various combinations of Cross-Entropy Loss, Focal Loss, and Dice Loss on model performance. As illustrated in Table 5, the results indicate that using all three in combination offers the greatest advantage. Specifically, Cross-Entropy Loss provides basic classification capabilities and is the most indispensable among three loss functions above. Focal Loss automatically focuses on those hard-to-segment boundary regions and ambiguous parts, which is particularly important at the edges of moving objects across videos. Dice Loss improves the integrity of the target contours by optimizing the overlap between regions.

Components			Performance
Versions	Text-D	S-Node	Hyper-G	$J\&F$
I	✗	✓	✓	**59.5**
II	✓	✗	✓	**59.7**
III	✓	✓	✗	**59.6**
IV	✓	✗	✗	**59.2**
V	✗	✓	✗	**59.0**
VI	✗	✗	✓	**59.3**
VII	✗	✗	✗	**58.9**
VIII	✓	✓	✓	**59.8**

Table 4: Ablation of various components in Ref-YouTube-VOS.

Type	CE Loss	Focal Loss	Dice Loss	Performance		
Impact	Class	Box	Mask	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
I	✓	✓	✗	58.5	57.3	59.7
II	✓	✗	✓	58.0	56.8	59.2
III	✓	✓	✓	59.8	58.4	61.1

Table 5: Combinatorial experiment of loss function.

4.4 Qualitative Results

Figure 3 shows a sequence of skydiving frames with different segmentation results. Text-depth collaborative integration: The system processes two different text expressions ("a person wearing a blue jumpsuit is being held by another with a parachute flying in the sky" and "a man behind another man in a harness"). The HTD method shows better consistency in segmenting the skydivers across frames while maintaining the spatial relationship described in the textual elements, especially in handling the complex overlapping of the two people.

5 Conclusion

In this paper, we have proposed a novel video object segmentation approach called Hyper-graph Text-Depth Collaborative Reasoning Video Object Segmentation (HTD), which achieves significant improvements over existing VOS techniques. By integrating multimodal information including appearance, location, text, and depth, HTD provides richer contextual information, thereby enhancing the accuracy and robustness of segmentation. Our main contributions are: First, we propose a progressive hyper-graph model incorporating multimodal information for hyper-edges between adjacent regions, enhancing target representation. Second, we improve reliability through multiple object proposals per frame. Finally, we introduce a collaborative reasoning mechanism that handles complex relationships while integrating various modalities. Experimental results on four challenging datasets demonstrate that HTD achieves excellent performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62176172, 61672364) and National Key Research and Development Program of China (No.2018YFA0701701).

References

[Antelmi *et al.*, 2023] Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. A survey on hypergraph representation learning. *ACM Computing Surveys*, 56(1):1–38, 2023.

[Botach *et al.*, 2022a] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022.

[Botach *et al.*, 2022b] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4985–4995, 2022.

[Chan *et al.*, 2018] T-H Hubert Chan, Anand Louis, Zhihao Gavin Tang, and Chenzi Zhang. Spectral properties of hypergraph laplacian and approximation algorithms. *Journal of the ACM (JACM)*, 65(3):1–48, 2018.

[Cheng *et al.*, 2024] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024.

[Cho *et al.*, 2024] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Dual prototype attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19238–19247, 2024.

[Ding *et al.*, 2025] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. In *European Conference on Computer Vision*, pages 215–233. Springer, 2025.

[Fan *et al.*, 2021] Haoyi Fan, Fengbin Zhang, Yuxuan Wei, Zuoyong Li, Changqing Zou, Yue Gao, and Qionghai Dai. Heterogeneous hypergraph variational autoencoder for link prediction. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4125–4138, 2021.

[Gavrilyuk *et al.*, 2018] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018.

[Han *et al.*, 2023] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *ICCV*, pages 13414–13423, October 2023.

[He and Ding, 2024] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, pages 13332–13341, 2024.

[Hu *et al.*, 2024] Xiao Hu, Basavaraj Hampiholi, Heiko Neumann, and Jochen Lang. Temporal context enhanced referring video object segmentation. In *WACV*, pages 5574–5583, 2024.

[Jhuang *et al.*, 2013a] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013.

[Jhuang *et al.*, 2013b] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013.

[Khoreva *et al.*, 2019] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, pages 123–141. Springer, 2019.

[Kim *et al.*, 2020] Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. Hypergraph attention networks for multimodal learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14581–14590, 2020.

[Li *et al.*, 2025] Wanyun Li, Pinxue Guo, Xinyu Zhou, Lingyi Hong, Yangji He, Xiangyu Zheng, Wei Zhang, and Wenqiang Zhang. Onevos: unifying video object segmentation with all-in-one transformer framework. In *European Conference on Computer Vision*, pages 20–40. Springer, 2025.

[Liang *et al.*, 2023] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *TPAMI*, 45(8):10055–10069, 2023.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022.

[Luo *et al.*, 2023] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *NeurIPS*, 36, 2023.

[Mao *et al.*, 2016] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[Miao *et al.*, 2023a] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *ICCV*, pages 920–930, October 2023.

[Miao *et al.*, 2023b] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023.

[Miao *et al.*, 2024] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Temporally consistent referring video object segmentation with hybrid memory. *TCSVT*, 2024.

[Peng *et al.*, 2025] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[Pont-Tuset *et al.*, 2017] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *CVPR*, 2017.

[Ranftl *et al.*, 2020] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 44(3):1623–1637, 2020.

[Seo *et al.*, 2020] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223. Springer, 2020.

[Tang *et al.*, 2023] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *ICCV*, pages 15466–15476, 2023.

[Wu *et al.*, 2022] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022.

[Wu *et al.*, 2023] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, pages 2761–2770, 2023.

[Wu *et al.*, 2024] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. *arXiv preprint arXiv:2411.10332*, 2024.

[Xu *et al.*, 2023] Yuanyou Xu, Zongxin Yang, and Yi Yang. Video object segmentation in panoptic wild scenes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Yang *et al.*, 2024] Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H.S. Torr. Language-aware vision transformer for referring segmentation. *TPAMI*, pages 1–18, 2024.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016.

[Zhu *et al.*, 2025] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *European Conference on Computer Vision*, pages 452–469. Springer, 2025.