MSTDiff: Multiscale-Aware Transformer Diffusion Network for Video Object Detection

Qiang Qi[1], Wenqi Shang[1], Xiao Wang[1]*, Yanjie Liang[2], Shuyuan Lin[3]

[1]School of Data Science, Qingdao University of Science and Technology, Qingdao, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]College of Cyber Security, Jinan University, Guangzhou, China
qiangq@qust.edu.cn, shangwenq@mails.qust.edu.cn, xiaowang@qust.edu.cn

Abstract

Video object detection is a fundamental yet challenging task in computer vision. Recently, DETR-based methods have gained prominence in this domain owing to their powerful global modeling capabilities. However, these methods are usually confronted with two key limitations: frame-agnostic initialization of object queries and scale-agnostic attention mechanisms, which hinder their capability to capture the appearance variations of dynamic objects and model temporal consistency across frames. To alleviate these limitations, we propose a multiscale-aware transformer diffusion network (MSTDiff), a novel framework designed for the video object detection task, including two technical improvements over existing methods. First, we design a diffusion-driven adaptive query module, which models the object query distribution through a diffusion process conditioned on input frames, enabling an adaptive and content-aware initialization of object queries. Second, we develop a multiscale-aware transformer encoder module, which combines multi-head convolutional units with attention mechanisms to enhance multiscale feature representations while preserving global dependence modeling. We conduct extensive experiments on the public ImageNet VID dataset, and the results demonstrate that our MSTDiff achieves 87.7% mAP with ResNet-101, outperforming most previous state-of-the-art video object detection methods.

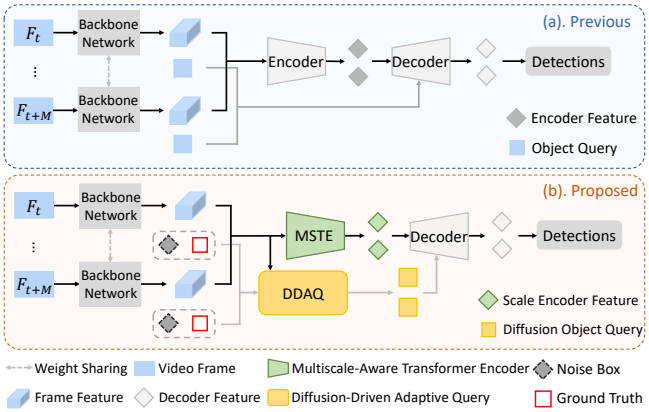

Figure 1: Comparison between (a) previous DETR-based video object detection methods and (b) the proposed MSTDiff. (a) Previous methods generally rely on the frame-agnostic initialization of object queries and scale-agnostic attention mechanisms, leading to suboptimal performance. (b) The proposed MSTDiff effectively alleviates the limitations of (a) by using the designed diffusion-driven adaptive query and multiscale-aware transformer encoder modules.

Introduction

Video object detection, a pivotal task in computer vision, aims to detect and locate objects in all frames of a given video and has received extensive research in recent years. It plays an important role in various applications, including robot navigation (Xu et al. 2023), autonomous driving (Cao et al. 2023), and human-computer interaction (Ni et al. 2023). Unlike static image object detection, video object detection needs to leverage temporal consistency and contextual information across frames to achieve accurate predictions. The key reason is attributed to complex and unpredictable appearance shifts caused by issues such as motion blur, occlusion, deformation, and pose changes. To tackle these issues, many methods attempt to exploit temporal information more effectively by aggregating features across frames. For example, MEGA (Chen et al. 2020) enhances video object detection by combining global and local temporal feature aggregation through a memory-enhanced framework. MAMBA (Sun et al. 2021a) employs a multi-level memory bank to aggregate temporal features across frames, effectively enhancing appearance representations. Although these methods improve detection performance, they typically rely on manually designed mechanisms for feature aggregation, such as predefined keyframe selection strategies and handcrafted memory update rules. These designs result in complicated pipelines and limited flexibility.

In light of this, DETR-based video object detection methods offer a new paradigm by leveraging spatial-temporal transformers with global attention and set-based prediction, enabling end-to-end detection without manually designed components and achieving impressive performance, as shown in Figure 1(a). TransVOD (Zhou et al. 2023) is one of the representative DETR-based video object methods, and it utilizes a spatial-temporal transformer to fuse object

*Corresponding Author.

queries across frames. PTSEFormer (Wang et al. 2022) designs a progressive temporal-spatial enhanced transformer to enhance spatial-temporal feature representations by aggregating information from multiple frames. ClipVID (Deng, Chen, and Wu 2023) adopts an identity-consistent transformer decoder to capture fine-grained temporal context and generate more comprehensive object representations.

Although DETR-based video object detection methods have achieved impressive performance on the benchmark datasets, they generally suffer from two fundamental limitations: 1) **Frame-agnostic initialization of object queries.** Most DETR-based video object detection methods adopt the randomly initialized learnable embeddings as initial object queries, which are inherently agnostic to the visual content of input frames, making them less effective in adapting to the diverse visual characteristics of objects. Since object appearances in videos usually change drastically across frames due to motion blur and object occlusion, the usage of frame-agnostic initial object queries may lead to suboptimal object association and inaccurate detection, especially in complex and dynamic scenes. 2) **Scale-agnostic attention mechanisms in transformers.** Most DETR-based video object detection methods apply uniform attention mechanisms across all feature scales without explicitly modeling scale variations, which limits their capability to accurately capture objects with varying sizes and dynamic motion patterns. Since objects in videos frequently undergo significant scale variations due to object movements, the usage of scale-agnostic attention mechanisms may struggle to maintain robust and consistent performance under such conditions.

To alleviate the aforementioned limitations, we propose a multiscale-aware transformer diffusion network (MSTDiff) for video object detection, as illustrated in Figure 1(b). Our MSTDiff is designed based on two main components: the diffusion-driven adaptive query (DDAQ) module and the multiscale-aware transformer encoder (MSTE) module. Specifically, the DDAQ module generates object queries through a learnable diffusion process conditioned on input frames. By sampling around the ground-truth boxes and injecting Gaussian noise into them, DDAQ produces diverse object queries that are adaptive to context-awareness variations under the guidance of input frames. The MSTE module combines attention mechanisms and multi-head convolutional blocks with varying kernel sizes, giving each head a distinct receptive field. This design enables the capture and fusion of multi-scale features across frames, which is crucial for detecting objects of varying sizes and rapid motion. We integrate the carefully designed DDAQ and MSTE modules into an end-to-end framework and evaluate it on the public ImageNet VID dataset. Our MSTDiff achieves substantial gains over prior methods, confirming its effectiveness in video object detection. Particularly, our MSTDiff achieves 87.7% mAP with the ResNet-101 backbone. In summary, the contributions of this work are presented as follows:

- We propose MSTDiff, a novel video object detection network combining diffusion with transformer models, and it achieves 87.7% mAP on the ImageNet VID dataset, showing strong performance and surpassing most existing video object detection methods.
- We propose a diffusion-driven adaptive query module, which leverages a denoising diffusion process to generate object queries, making them context-aware under the guidance of input frames.
- We propose a multiscale-aware transformer encoder module that seamlessly integrates multi-head convolutional units with attention mechanisms, enabling effective multi-scale representations while preserving strong global context modeling capability.

Related Work

Video Object Detection

Video object detection aims to address the appearance deterioration issue in videos, such as motion blur and object occlusion. To tackle this issue, many studies focus on improving the detection of the current frame by leveraging the temporal information across frames. For example, SELSA (Wu et al. 2019) provides a new perspective for video object detection by performing cross-frame semantic aggregation on high-level features and introducing attention mechanisms to enhance feature fusion. PTSEFormer (Wang et al. 2022) proposes a progressive strategy to enhance temporal and spatial information by establishing attention between the target frame and support frames and leveraging a spatial transformation perception module to transmit positional information between regions. CETR (An et al. 2024) introduces a novel framework that improves current frame detection by leveraging class-wise memory and classification-based sampling to selectively utilize relevant contextual information across frames, thereby enhancing robustness against motion blur and occlusion. DGC-Net (Qi et al. 2025) adopts a dynamic graph contrastive framework to model intra-class object relationships across frames, where a graph structure is constructed to capture object interactions and propagate temporal context, facilitating more accurate object localization and classification.

By contrast, our MSTDiff introduces a diffusion model to reconstruct object queries from a new perspective and incorporates the scale information into the transformer encoder to effectively model the spatial-temporal feature representations of objects.

Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) were proposed as a novel class of generative models based on iterative denoising processes inspired by nonequilibrium thermodynamics. Building on this theoretical foundation, denoising diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) demonstrated their strengths in the image generation domain, showing impressive sample quality and flexible conditioning capabilities. Following their success in image generation, diffusion models have been extended to other domains, including text generation (Lin et al. 2023b), video generation (Blattmann et al. 2023; Ho et al. 2022), and medical image analysis (Wu et al. 2024). Their ability to model complex distributions and incorporate various forms of conditioning makes them particularly suitable for structured pre-

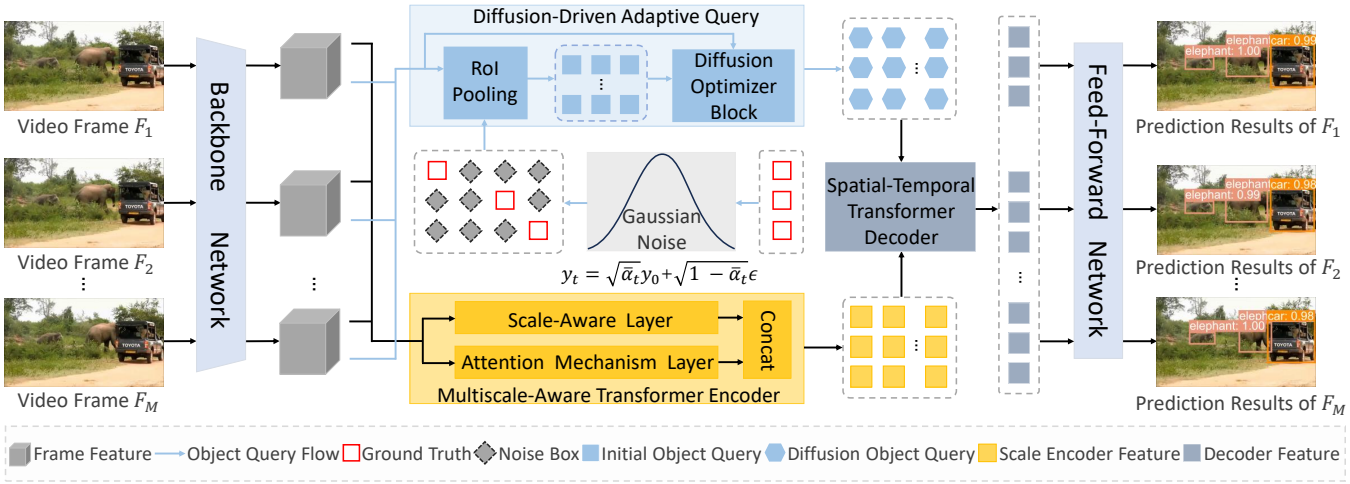

Figure 2: Overall framework of the proposed multiscale-aware transformer diffusion network (MSTDiff).

diction tasks. Recently, diffusion-based methods have also been extensively explored in the field of image object detection. For instance, DiffusionDet (Chen et al. 2023) formulates object detection as a denoising process from random boxes to object boxes, enabling the dynamic number of boxes and iterative refinement. Motivated by the effectiveness of diffusion-based detection in static images, our MSTDiff generates object queries through a learnable diffusion process conditioned on the input frames, enabling an adaptive and content-aware initialization of object queries.

Methodology

Preliminaries: Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) formulate generative modeling as a latent variable process, consisting of a forward noising procedure and a learnable reverse denoising trajectory to reconstruct the data distribution. Formally, given a data sample $y_0 \sim q(y_0)$, the forward diffusion process progressively adds Gaussian noise at each time step $t \in \{1, 2, \cdots, T\}$ according to a predefined variance schedule β_t. This process is defined as:

$$q(y_t \mid y_{t-1}) := \mathcal{N}(y_t \mid \sqrt{1-\beta_t} y_{t-1}, \beta_t \mathbf{I}). \quad (1)$$

To obtain y_t directly from y_0, one can sample a Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and compute:

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad (2)$$

where $\bar{\alpha}_t = \prod_{k=1}^{t}(1-\beta_k)$. During training, a neural network $\epsilon_\theta(y_t, t)$ is trained to predict the added Gaussian noise ϵ given the noisy input y_t and the timestep t, typically using the following simplified objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{y_0, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(y_t, t)\|^2\right]. \quad (3)$$

At the inference stage, the diffusion models start from pure Gaussian noise $y_T \sim \mathcal{N}(0, \mathbf{I})$, and iteratively apply the learnable reverse process $p_\theta(y_{t-1} \mid y_t)$ to generate a clean sample y_0.

Framework Overview: MSTDiff

The overall framework of the proposed multiscale-aware transformer diffusion network (MSTDiff) is shown in Figure 2, which takes multiple frames as inputs and outputs the prediction results for each corresponding frame. Specifically, each frame in the video sequence is independently passed through a shared backbone to extract the frame features $f_m \in \mathbb{R}^{C \times H \times W}$, where m denotes the frame index, with C, H, and W representing the number of dimensions, height, and width, respectively. These frame features serve as the foundation for subsequent spatio-temporal modeling. Building on these features, our MSTDiff introduces two core modules. The first one is a diffusion-driven adaptive query module, which models the object query distribution through a diffusion process conditioned on input frames, generating diffusion object queries. The second one is a multiscale-aware transformer encoder module, which integrates multi-head convolutional operations and attention mechanisms to capture multi-scale features while preserving global dependency modeling, outputting scale encoder features. After that, the diffusion object queries and scale encoder features are fed into a spatial-temporal transformer decoder module to conduct query-feature interaction. Finally, a shared feed-forward network is employed as the detection head to produce the final prediction results.

Diffusion-Driven Adaptive Query Module

Building upon the advances in diffusion-based object detection (Chen et al. 2023; Zhang et al. 2025; Roh and Chung 2023), we introduce the diffusion-driven adaptive query (DDAQ) module, with the goal of using frame features to adaptively generate object queries through the diffusion process. Specifically, we initialize the data samples of m-th frame with a set of noisy boxes $\mathbf{y}_T = \mathbf{Q}_m^0$. These noise boxes are sampled from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and serve as the starting point of the reverse diffusion process. To obtain initial object queries that reflect the semantic content of each frame, the noisy boxes $\mathbf{Q}_m^0$

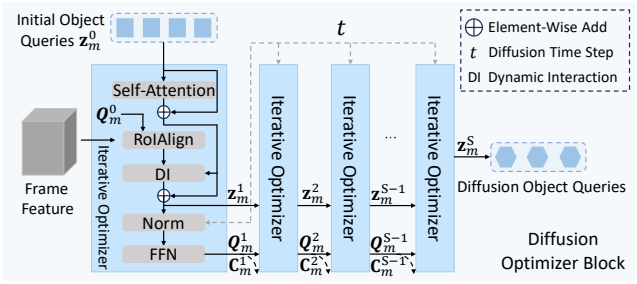

Figure 3: The schematic of the proposed diffusion optimizer block (DOB).

are used to extract localized features from the frame features f_m. The resulting initial object queries $\mathbf{z}_m^0$ for the m-th frame are computed using a region-level feature extractor:

$$\mathbf{z}_m^0 = \text{AvgPool}\Big(\text{RoIAlign}(\mathbf{Q}_m^0, f_m)\Big), \tag{4}$$

where $\mathbf{z}_m^0 \in \mathbb{R}^{N\times D}$, N is the number of initial object queries, and D is the feature dimension. RoIAlign$(\cdot,\cdot)$ denotes the region feature extraction function. AvgPool$(\cdot)$ indicates the global average pooling operation over the extracted region features. As the initial object queries are derived from arbitrary regions, they typically lack rich semantic and localization cues. To address this limitation, we design a diffusion optimizer block (DOB), as illustrated in Figure 3, which is composed of multiple iterative optimizers that progressively enrich instance-level representations. Through multiple stages, the iterative optimizer updates each object query by aggregating contextual information from other object queries and attending to the regions within its own predicted boxes, thereby enabling richer representations. At each stage of the iterative optimizer, the object queries are progressively refined through the self-attention mechanism and instance-level interactions, conditioned on the diffusion time steps to enhance their representation capability. Formally, the multi-stage multi-head self-attention applied to the object queries at the stage s can be computed as:

$$\text{MHSAttn}(\mathbf{z}_m^s) = \text{Concat}(\text{head}_1, \cdots, \text{head}_O)\mathbf{W}^J, \tag{5}$$

$$\text{head}_o = \text{Softmax}\left(\frac{(\mathbf{z}_m^s\mathbf{W}_o^Q)(\mathbf{z}_m^s\mathbf{W}_o^K)^\top}{\sqrt{D_k}}\right)(\mathbf{z}_m^s\mathbf{W}_o^V), \tag{6}$$

where $\mathbf{z}_m^s \in \mathbb{R}^{N\times D}$ denotes the object queries of the m-th frame at the s-th stage. O is the number of attention heads. $\mathbf{W}_o^Q$, $\mathbf{W}_o^K \in \mathbb{R}^{D\times D_k}$, $\mathbf{W}_o^V \in \mathbb{R}^{D\times D_r}$, $\mathbf{W}^J \in \mathbb{R}^{OD_r\times D}$ are the learnable projection matrices. The attention weights are computed by the scaled dot-product between the projected queries and keys, and the output is the weighted sum over the projected values. Each attention head independently projects the input into query, key, and value vectors of the dimensions D_k, D_k, and D_r, respectively. The results of all heads are concatenated and projected to the final representation via $\mathbf{W}^J$. After that, a RoIAlign layer is applied on the predicted boxes $\mathbf{Q}_m^s$ and the frame features f_m to extract fine-grained region features. To integrate the region features into the object query representation, a specialized operation known as dynamic interaction (Sun et al. 2021b) is employed. In this step, each object query attends to the region features extracted based on its previously predicted boxes. The region features are modulated by parameters dynamically generated from each object query through fully connected layers. Additionally, the diffusion time step t is encoded through linear transformations to produce time-dependent embeddings, which are used to normalize the object queries and enable a multistep reverse process as in the DDIM (Song, Meng, and Ermon 2020).

To provide a clear and concise representation of the update process, we formulate the object query refinement at the s-th stage as:

$$\mathbf{z}_m^s = \mathcal{R}^s\left(\mathbf{z}_m^{s-1}, \mathbf{Q}_m^{s-1}, f_m, t\right), \quad s \in \{1, \cdots, S\}, \tag{7}$$

where $\mathbf{z}_m^s \in \mathbb{R}^{N\times D}$ denotes the object query of the m-th frame at the s stage, $\mathcal{R}^s(\cdot,\cdot,\cdot,\cdot)$ denotes the iterative refinement operation at the stage s, $\mathbf{Q}_m^{s-1}$ is the predicted boxes of the m-th frame at the stage $s-1$, f_m represents the frame features of the m-th frame, and t indicates the current diffusion time step.

At the end of each iterative optimizer stage, the optimized object queries $\mathbf{z}_m^s$ are processed by a feed-forward network (FFN) to produce the corresponding predictions, including the predicted boxes $\mathbf{Q}_m^s$ and classification scores $\mathbf{C}_m^s$. The predicted boxes $\mathbf{Q}_m^s$ are then used in the next iterative optimizer stage as spatial references to extract new region features. Finally, the object queries in the last stage, denoted as diffusion object queries $\mathbf{z}_m^S$, are fed into the spatial-temporal transformer decoder module.

Multiscale-Aware Transformer Encoder Module

The schematic of the proposed multiscale-aware transformer encoder (MSTE) module is illustrated in Figure 4(a). The goal of MSTE is to combine multi-head convolution units with attention mechanisms to obtain effective multi-scale features while retaining strong global feature representations. For the scale layer, the frame features are first fed into a scale fusion block (SFB), which is the core of MSTE, with its schematic shown in Figure 4(b). Within SFB, the input features are first projected by a linear layer and then passed through the mixed depth-wise convolution (MDWC) operation to capture multi-scale spatial patterns. MDWC divides the input channels into multiple heads and applies distinct depth-wise separable convolutions (Chollet 2017; Lu, Zhang, and Wang 2021; Lin et al. 2023a) to each head. In detail, the frame features $x = f_m \in \mathbb{R}^{H\times W\times C}$ serve as the input and are split into A heads along the channel dimension, denoted as $x = [x_1, x_2, \cdots, x_A]$, and each head is processed with a convolution of kernel size $p_i \in \{3, 5, \cdots, P\}$, where p_i increases by 2 with each head. Particularly, the MDWC operation is formulated as:

$$\text{MDWC}(x) = \text{Concat}_{\text{channel}}\left(\text{DW}_{p_i}(x_i)\right)_{i=1}^{A}, \tag{8}$$

where $\text{DW}_{p_i}(x_i)$ denotes the depth-wise convolution applied to the feature of the i-th head x_i with the kernel size $p_i \times p_i$. This multi-head design allows each head to focus on

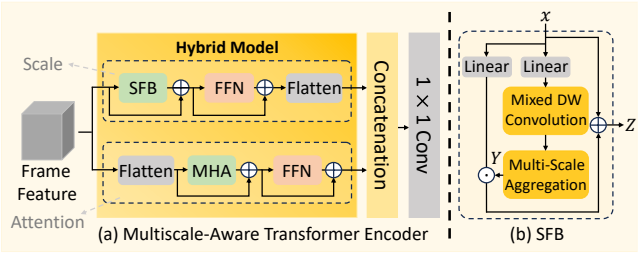

Figure 4: The schematic of the proposed multiscale-aware transformer encoder (MSTE) and scale fusion block (SFB).

spatial patterns at different scales, capturing the fine-grained features. Compared with single-head convolutions, MDWC improves the model's ability to attend to target regions while suppressing background noises, and retains object details even in deeper stages of the network, thereby enhancing the feature representation capability with only moderate computational cost. The process of the multi-scale aggregation (MSA) block is formulated as:

$$U_l = \mathbf{W}_{\text{intra}} \left([G_1^l, G_2^l, \cdots, G_A^l] \right), \tag{9}$$

$$Y = \mathbf{W}_{\text{inter}} \left([U_1, U_2, \cdots, U_E] \right), \tag{10}$$

where $\mathbf{W}_{\text{inter}}$ and $\mathbf{W}_{\text{intra}}$ denote the weight matrices of the point-wise convolution. Y denotes the refined features after being processed by MSA. $G_i^l = \text{DW}_{p_i \times p_i}(x_i^l)$ denotes the features obtained by applying a depth-wise convolution to the l-th channel in the i-th head. Let $i \in \{1, 2, \cdots, A\}$ and $l \in \{1, 2, \cdots, E\}$, in which A is the number of heads and $E = C/A$ is the number of groups. $G_i \in \mathbb{R}^{H \times W \times E}$ represents the i-th depth-wise convolutional head. Next, these refined features are used to dynamically modulate the value tensor M via element-wise multiplication. For the input features x, the output features of SFB, denoted as Z, can be formulated as:

$$Z = Y \odot M \oplus x, \tag{11}$$

$$M = \mathbf{W}_m x, \tag{12}$$

where $\odot$ denotes the element-wise multiplication. The features Y can be adaptively adjusted based on the input content, enabling the dynamic modulation. Unlike the self-attention mechanism, which requires computing an $A \times A$ attention matrix, this modulator preserves the channel dimension. As a result, after performing element-wise multiplication, it achieves targeted modulation across both spatial and channel dimensions. This property not only enhances the feature representation capability but also improves memory efficiency when processing high-resolution frames. Subsequently, a feed-forward network (FFN), a residual connection, and a flattening operation are applied to further refine and reshape the features Z, which are then used as the output of the scale layer.

For the attention layer, the input features are first flattened and then processed by a multi-head self-attention block to enhance the modeling of long-range dependencies. The calculation of the multi-head self-attention is presented in Equations (5) and (6), after which a feed-forward network (FFN) and a residual connection are applied. Finally, the outputs of the two layers are concatenated and processed by a 1×1 convolution to obtain the final outputs of MSTE.

Experiments

Experimental Settings

Dataset. To evaluate the effectiveness of our MSTDiff, we conduct extensive experiments on the ImageNet VID (Russakovsky et al. 2015) dataset. This dataset includes 3862 training videos and 555 validation videos, annotated with bounding boxes across 30 diverse object categories. Following existing video object detection methods (Qi et al. 2025; An et al. 2024; Qiu et al. 2024; Roh and Chung 2023; Deng, Chen, and Wu 2023), we adopt mean average precision (mAP) as the evaluation metric.

Implementation Details. Our MSTDiff is trained on 4 24GB RTX-4090 GPUs with a batch size of 4. Following the widely adopted implementation protocols in previous methods (Qi, Yan, and Wang 2023; Deng, Chen, and Wu 2023; Roh and Chung 2023), we use ResNet-101 (He et al. 2016), ResNeXt-101 (Xie et al. 2017), and Swin-Base (Liu et al. 2021) as the backbone networks for evaluation. Deformable DETR (Zhu et al. 2020) is adopted as the baseline, and AdamW (Loshchilov and Hutter 2017) is adopted as the optimizer with a weight decay of 10^{-4}. The learning rate is set to 2×10^{-4} for the first 80K iterations and 2×10^{-5} for the last 40K iterations. The transformer weights are initialized using Xavier initialization (Glorot and Bengio 2010), while the backbone networks are initialized from an ImageNet-pretrained model (Deng et al. 2009). By default, we adopt $M = 20$ frames as inputs and set the number of diffusion object queries to 100. During training, all frames are augmented using the same strategies (Wu et al. 2019), including random horizontal flipping and random resizing. Each input frame is resized so that its shorter side is at least 600 pixels and its longer side is limited to 1000 pixels.

Comparisons with State-of-the-Art

We compare the proposed MSTDiff with a broad range of state-of-the-art video object detection methods on the ImageNet VID dataset, as illustrated in Table 1. Our MSTDiff demonstrates consistent and competitive performance under different backbone configurations. When using ResNet-101 as the backbone and Deformable DETR as the baseline, our MSTDiff achieves an mAP of 87.7%, surpassing strong methods such as HyMAT (Moorthy et al. 2025) and CDANet (Qi, Yan, and Wang 2023). In particular, our MSTDiff outperforms HyMAT, a recently published transformer-based method, by 1.0% mAP, despite using the same backbone and a similar detection head. Compared to TransVOD++ (Zhou et al. 2023), which also adopts Deformable DETR, our MSTDiff achieves an improvement of 5.7% mAP, highlighting the superiority of our diffusion-driven adaptive query design over the conventional object query mechanism. We further validate the performance of our MSTDiff by equipping it with stronger backbones ResNeXt-101 and Swin-Base. The results show that our MSTDiff also displays clear advantages. Specifically, our

Method	Venue	Backbone	Base Detector	mAP(%)
MEGA (Chen et al. 2020)	CVPR	ResNet-101	Faster R-CNN	82.9
LSTS (Jiang et al. 2020)	ECCV	ResNet-101	R-FCN	77.2
HVRNet (Han et al. 2020)	ECCV	ResNet-101	Faster R-CNN	83.2
DSFNet (Lin et al. 2020)	ACM MM	ResNet-101	Faster R-CNN	84.1
MAMBA (Sun et al. 2021a)	AAAI	ResNet-101	Faster R-CNN	84.6
Tf-blender (Cui et al. 2021)	ICCV	ResNet-101	Faster R-CNN	83.8
QueryProp (He et al. 2022)	AAAI	ResNet-101	Sparse R-CNN	82.3
EOVOD (Sun et al. 2022)	ECCV	ResNet-101	Faster R-CNN	79.8
TransVOD++ (Zhou et al. 2023)	ACM MM	ResNet-101	Deformable DETR	82.0
MSTF (Xu et al. 2022)	TCSVT	ResNet-101	Faster R-CNN	83.3
ClipVID (Deng, Chen, and Wu 2023)	ICCV	ResNet-101	DETR	84.7
GMLCN (Han and Yin 2023)	TMM	ResNet-101	Faster R-CNN	78.6
CDANet (Qi, Yan, and Wang 2023)	TMM	ResNet-101	Faster R-CNN	85.4
CETR (An et al. 2024)	AAAI	ResNet-101	DAB-DETR	79.6
PDMAN (Qiu et al. 2024)	ICASSP	ResNet-101	Faster R-CNN	83.9
HyMAT (Moorthy et al. 2025)	EAAI	ResNet-101	DETR	86.7
MSTDiff (ours)	-	ResNet-101	Deformable DETR	**87.7**
HVRNet (Han et al. 2020)	ECCV	ResNeXt-101	Faster R-CNN	84.8
ClipVID (Deng, Chen, and Wu 2023)	ICCV	ResNeXt-101	DETR	85.8
DGC-Net (Qi et al. 2025)	TIP	ResNeXt-101	Faster R-CNN	87.3
MSTDiff (ours)	-	ResNeXt-101	Deformable DETR	**88.5**
SwinVid (Maharek et al. 2024)	CSSE	Swin-Base	Faster R-CNN	84.3
TransVOD++ (Zhou et al. 2023)	TPAMI	Swin-Base	Deformable DETR	90.0
STPN (Sun et al. 2023)	ICCV	Swin-Base	SELSA	90.6
MSTDiff (ours)	-	Swin-Base	Deformable DETR	**91.5**

Table 1: Performance comparison between our MSTDiff and some existing video object detection methods on the ImageNet VID validation set.

MSTDiff achieves 88.5% mAP with the ResNeXt-101 backbone, outperforming other methods that use the same backbone. When using the Swin-Base backbone, our MSTDiff outperforms SwinVid (Maharek et al. 2024) by a margin of 7.2% mAP. This improvement is attributed to the utilization of our specifically designed diffusion-driven adaptive query and multiscale-aware transformer encoder modules, both of which enhance spatial-temporal feature representations of objects. All these results in Table 1 demonstrate the effectiveness of our MSTDiff.

Ablation Study

Effect of Each Component in MSTDiff. Table 2 summarizes the effect of different components in MSTDiff on the ImageNet VID dataset. Starting from model A, the baseline detector Deformable DETR with ResNet-101, achieves 78.5% mAP. Model B introduces the diffusion-driven adaptive query (DDAQ) module into model A, which brings a notable mAP improvement of 3.4%, reaching 81.9%, demonstrating the effectiveness of our DDAQ module. Model C incorporates the multiscale-aware transformer encoder (MSTE) module into model A, which improves the mAP from 78.5% to 86.1%, highlighting the effectiveness and benefits of our MSTE module. Finally, model D integrates both DDAQ and MSTE modules into model A, resulting in a final mAP of 87.7%. These consistent improvements validate the individual effectiveness and synergistic contributions of the proposed components in MSTDiff.

Model	Baseline	DDAQ	MSTE	mAP (%)
A	✓			78.5
B	✓	✓		81.9
C	✓		✓	86.1
D	✓	✓	✓	**87.7**

Table 2: Ablation studies of the proposed components.

Effect of the Number of Iterative Optimization Steps. To fully evaluate the effect of the number of iterative optimization steps within the diffusion-driven adaptive query module on the overall detection accuracy, we conduct a series of experiments by varying the number of steps s. The results are summarized in Table 3. From the table, we can observe consistent improvements in detection accuracy as the number of steps s increases from 1 to 4, with the mAP progressively rising from 86.7% to 87.7%. This highlights the benefit of progressively refining object queries through the iterative diffusion-based optimization process, which helps our MSTDiff to better capture object-specific features in a coarse-to-fine manner. However, increasing the number of steps to $s = 5$ does not yield additional mAP gain. This detection accuracy saturation may be attributed to the over-smoothing of the object queries or potential optimization redundancy, where too many steps do not bring new discriminative information. Based on this observation, we empirically set $s = 4$

Step s	mAP (%)	ΔmAP (%)
1	86.7	–
2	87.1	+0.4
3	87.5	+0.4
4	**87.7**	+0.2
5	**87.7**	+0.0

Table 3: Effect of the number of iterative optimization steps s in the diffusion-driven adaptive query module. ΔmAP denotes the improvement compared to the previous step.

Noisy Boxes	100	200	300	350	400
mAP (%)	85.4	86.9	**87.7**	87.6	87.3

Table 4: Effect of the number of noisy boxes. Performance peaks at 300 noisy boxes, with performance degradation observed at more noisy boxes.

as the optimal configuration in all experiments.

Effect of the Number of Noisy Boxes. To fully analyze the effect of the number of noisy boxes on detection accuracy, we vary the number of noisy boxes and report the experimental results in Table 4. From the results, we can observe that as the number of noisy boxes increases from 100 to 300, the mAP consistently improves from 85.4% to a peak value of 87.7%. However, when the number of noisy boxes exceeds 300, the mAP begins to degrade, possibly due to the redundancy and noisy box accumulation that impair the object query optimization process. Specifically, when the number of noisy boxes is increased from 300 to 400, the mAP degrades from 87.7% mAP to 87.3 % mAP. All these results in Table 4 demonstrate that using an appropriate number of noisy boxes to guide the diffusion-based object query denoising process can enhance the robustness of our MSTDiff by encouraging it to learn more discriminative features. Therefore, we set the number of noisy boxes to 300 as our final design in all experiments.

Effect of Different Stacking Strategies. To fully explore the effect of different stacking strategies on detection accuracy, we perform a series of ablation experiments, and the experimental results are illustrated in Table 5. In detail, the hybrid stacking strategy that employs parallel SFB (scale fusion block) and MHSA (multi-head self-attention) achieves the best result with an accuracy of 87.7% mAP, which outperforms using only MHSA or SFB individually. This suggests that explicitly modeling the transition from local to global dependencies in a parallel manner is beneficial for improving detection accuracy. In contrast, placing SFB before MHSA yields inferior detection accuracy of 87.2% mAP, indicating that the sequential stacking strategy may not fully exploit the complementary advantages of both SFB and MHSA. Notably, the model using only MHSA still maintains a competitive detection accuracy of 86.7% mAP, further highlighting the effectiveness of global attention mechanisms in capturing long-range dependencies. All

Stacking Strategy	Hybrid	mAP (%)
SFB Only	✗	80.9
MHSA Only	✗	86.7
Parallel: SFB and MHSA	✓	**87.7**
Sequential: SFB and MHSA	✓	87.2

Table 5: Comparison of different stacking strategies. We see that the parallel stacking strategy achieves the best result.

Number of Heads	2	3	4	5	6
mAP (%)	86.7	87.2	87.5	**87.7**	87.5

Table 6: Effect of the number of heads in MDWC. We vary the number of heads from 2 to 6, and observe that the best accuracy is achieved when using 5 heads.

the results consistently demonstrate that adopting a parallel hybrid stacking strategy is optimal, validating the effectiveness of our MSTE module design.

Effect of the Number of Heads in MDWC. To investigate the effect of the number of heads in the mixed depth-wise convolution (MDWC) block, we perform a series of ablation experiments by varying the number of heads employed in MDWC. These heads are designed to capture spatial features at different scales. As illustrated in Table 6, increasing the number of heads improves the detection accuracy in the beginning, which demonstrates that multiple heads can help our MSTDiff learn richer and more diverse feature representations. Nevertheless, when the number of heads increases beyond a certain value, the detection accuracy begins to slightly diminish. This may stem from over-segmentation of feature channels, which limits the capability of each head to capture comprehensive information. Among all experimental configurations, the best accuracy is achieved when using 5 heads, reaching a maximum mAP of 87.7%. Therefore, we set the number of heads to 5 in all experiments.

Conclusion

In this paper, we propose a novel multiscale-aware transformer diffusion network (MSTDiff) tailored for video object detection. Our MSTDiff effectively alleviates two critical limitations of existing DETR-based video object detection methods: frame-agnostic initialization of object queries and scale-agnostic attention mechanisms. Specifically, we introduce a diffusion-driven adaptive query module, which enables dynamic and content-aware initialization of object queries through a diffusion process conditioned on input frames. Additionally, we design a multiscale-aware transformer encoder that integrates multi-head convolutional units with attention mechanisms to enhance scale perception of features while preserving global contextual modeling. Extensive experiments on the ImageNet VID dataset demonstrate that our MSTDiff delivers superior detection performance over state-of-the-art methods: 87.7% mAP with the ResNet-101 backbone, validating the effectiveness of our framework design.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62501343, 62476112, and 62202249; in part by the Natural Science Foundation of Shandong Province under Grants ZR2024QF294 and ZR2025QC1576; in part by the Guangdong Basic and Applied Basic Research Foundation under Grants 2024A1515011740 and 2025A1515010181; and in part by the Natural Science Foundation of Qingdao City under Grant 25-1-1-101-zyyd-jch.

References

An, S.; Park, S.; Kim, G.; Baek, J.; Lee, B.; and Kim, S. 2024. Context enhanced transformer for single image object detection in video data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 682–690.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Cao, Y.; Li, C.; Peng, Y.; and Ru, H. 2023. MCS-YOLO: A multiscale object detection method for autonomous driving road environment recognition. *IEEE Access*, 22342–22354.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19830–19843.

Chen, Y.; Cao, Y.; Hu, H.; and Wang, L. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10337–10346.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Cui, Y.; Yan, L.; Cao, Z.; and Liu, D. 2021. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8138–8147.

Deng, C.; Chen, D.; and Wu, Q. 2023. Identity-consistent aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13434–13444.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 249–256.

Han, L.; and Yin, Z. 2023. Global memory and local continuity for video object detection. *IEEE Transactions on Multimedia*, 3681–3693.

Han, M.; Wang, Y.; Chang, X.; and Qiao, Y. 2020. Mining inter-video proposal relations for video object detection. In *Proceedings of the European Conference on Computer Vision*, 431–446.

He, F.; Gao, N.; Jia, J.; Zhao, X.; and Huang, K. 2022. Queryprop: Object query propagation for high-performance video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 834–842.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the Conference on Neural Information Processing Systems*, 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. In *Proceedings of the Conference on Neural Information Processing Systems*, 8633–8646.

Jiang, Z.; Liu, Y.; Yang, C.; Liu, J.; Gao, P.; Zhang, Q.; Xiang, S.; and Pan, C. 2020. Learning where to focus for efficient video object detection. In *Proceedings of the European Conference on Computer Vision*, 18–34.

Lin, L.; Chen, H.; Zhang, H.; Liang, J.; Li, Y.; Shan, Y.; and Wang, H. 2020. Dual semantic fusion network for video object detection. In *Proceedings of the ACM International Conference on Multimedia*, 1855–1863.

Lin, W.; Wu, Z.; Chen, J.; Huang, J.; and Jin, L. 2023a. Scale-aware modulation meet transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6015–6026.

Lin, Z.; Gong, Y.; Shen, Y.; Wu, T.; Fan, Z.; Lin, C.; Duan, N.; and Chen, W. 2023b. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *Proceedings of the International Conference on Machine Learning*, 21051–21064.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, G.; Zhang, W.; and Wang, Z. 2021. Optimizing depthwise separable convolution operations on gpus. *IEEE Transactions on Parallel and Distributed Systems*, 70–87.

Maharek, A.; Abozeid, A.; Orban, R.; and ElDahshan, K. 2024. SwinVid: Enhancing Video Object Detection Using Swin Transformer. *Computer Systems Science and Engineering*, 305–320.

Moorthy, S.; KS, S. S.; Arthanari, S.; Jeong, J. H.; and Joo, Y. H. 2025. Hybrid multi-attention transformer for robust video object detection. *Engineering Applications of Artificial Intelligence*, 109606.

Ni, Z.; Mascaró, E. V.; Ahn, H.; and Lee, D. 2023. Human–object interaction prediction in videos through gaze following. *Computer Vision and Image Understanding*, 103741.

Qi, Q.; Wang, H.; Yan, Y.; and Li, X. 2025. DGC-Net: Dynamic Graph Contrastive Network for Video Object Detection. *IEEE Transactions on Image Processing*, 2269–2284.

Qi, Q.; Yan, Y.; and Wang, H. 2023. Class-aware dual-supervised aggregation network for video object detection. *IEEE Transactions on Multimedia*, 2109–2123.

Qiu, Z.; Qi, Q.; Lu, Y.; Yan, Y.; and Wang, H. 2024. Proposal Distillation of Multi-Modal Feature Aggregation Network for Video Object Detection. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 3895–3899.

Roh, S.-D.; and Chung, K.-S. 2023. DiffusionVID: Denoising Object Boxes With Spatio–Temporal Conditioning for Video Object Detection. *IEEE Access*, 121434–121444.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 211–252.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2256–2265.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, G.; Hua, Y.; Hu, G.; and Robertson, N. 2021a. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2620–2627.

Sun, G.; Hua, Y.; Hu, G.; and Robertson, N. 2022. Efficient one-stage video object detection by exploiting temporal consistency. In *Proceedings of the European Conference on Computer Vision*, 1–16.

Sun, G.; Wang, C.; Zhang, Z.; Deng, J.; Zafeiriou, S.; and Hua, Y. 2023. Spatio-temporal prompting network for robust video feature extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13587–13597.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021b. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14454–14463.

Wang, H.; Tang, J.; Liu, X.; Guan, S.; Xie, R.; and Song, L. 2022. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In *Proceedings of the European Conference on Computer Vision*, 732–747.

Wu, H.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9217–9225.

Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; and Xu, Y. 2024. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Proceedings of the Conference on Medical Imaging with Deep Learning*, 1623–1639.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Xu, C.; Zhang, J.; Wang, M.; Tian, G.; and Liu, Y. 2022. Multilevel spatial-temporal feature aggregation for video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 7809–7820.

Xu, Z.; Zhan, X.; Xiu, Y.; Suzuki, C.; and Shimada, K. 2023. Onboard dynamic-object detection and tracking for autonomous robot navigation with rgb-d camera. *IEEE Robotics and Automation Letters*, 651–658.

Zhang, M.; Wu, J.; Ren, Y.; Yang, J.; Li, M.; and Ma, A. J. 2025. Diffusionengine: Diffusion model is scalable data engine for object detection. *Pattern Recognition*, 112141.

Zhou, Q.; Li, X.; He, L.; Yang, Y.; Cheng, G.; Tong, Y.; Ma, L.; and Tao, D. 2023. TransVOD: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7853–7869.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.