Referred by Multi-Modality: A Unified Temporal Transformer for Video Object Segmentation

Shilin Yan[1,2*], **Renrui Zhang**[2,3*], **Ziyu Guo**[2*], **Wenchao Chen**[1]
Wei Zhang[1†], **Hongyang Li**[2†], **Yu Qiao**[2], **Zhongjiang He**[4], **Peng Gao**[2†]

[1]School of Computer Science, Fudan University
[2]Shanghai Artificial Intelligence Laboratory [3]CUHK MMLab
[4]China Telecom Corporation Ltd. Data&AI Technology Company
{tattoo.ysl}@gmail.com, {zhangrenrui, gaopeng}@pjlab.org.cn

Abstract

Recently, video object segmentation (VOS) referred by multi-modal signals, e.g., language and audio, has evoked increasing attention in both industry and academia. It is challenging for exploring the semantic alignment within modalities and the visual correspondence across frames. However, existing methods adopt separate network architectures for different modalities, and neglect the inter-frame temporal interaction with references. In this paper, we propose **MUTR**, a **M**ulti-modal **U**nified **T**emporal transformer for **R**eferring video object segmentation. With a unified framework for the first time, MUTR adopts a DETR-style transformer and is capable of segmenting video objects designated by either text or audio reference. Specifically, we introduce two strategies to fully explore the temporal relations between videos and multi-modal signals. Firstly, for low-level temporal aggregation before the transformer, we enable the multi-modal references to capture multi-scale visual cues from consecutive video frames. This effectively endows the text or audio signals with temporal knowledge and boosts the semantic alignment between modalities. Secondly, for high-level temporal interaction after the transformer, we conduct inter-frame feature communication for different object embeddings, contributing to better object-wise correspondence for tracking along the video. On Ref-YouTube-VOS and AVSBench datasets with respective text and audio references, MUTR achieves **+4.2%** and **+4.2%** $\mathcal{J}\&\mathcal{F}$ improvements to *state-of-the-art* methods, demonstrating our significance for unified multi-modal VOS. Code is released at https://github.com/OpenGVLab/MUTR.

1 Introduction

Multi-modal video object segmentation (VOS) aims to track and segment particular object instances across the video sequence referred by a given multi-modal signal, including referring video object segmentation (RVOS) with language reference, and audio-visual video object segmentation (AV-VOS) with audio reference. Different from the vanilla VOS with only visual information, the multi-modal VOS is more challenging and in urgent demand, which requires a comprehensive understanding among different modalities and their temporal correspondence across frames.

There exist two main challenges in multi-modal VOS. Firstly, it requires to not only explore the rich spatial-temporal consistency in a video, but also align the multi-modal semantics among image,

[*]Equal contribution. [†] Corresponding author.

Preprint. Under review.

language and audio. Current approaches mainly focus on the visual-language or visual-audio modal fusion within independent frames, simply by cross-modal attention [6, 28, 52] or dynamic convolutions [43] for feature interaction. This, however, neglects the multi-modal temporal information across frames, which is significant for consistent object segmentation and tracking along the video. Secondly, for the given references of two modalities, language and audio, existing works adopt different architecture designs and training strategies to separately tackle their modal-specific characteristics. Therefore, a powerful and unified framework for multi-modal VOS still remains an open question.

To address these challenges, we propose **MUTR**, a **M**ulti-modal **U**nified **T**emporal transformer for **R**eferring video object segmentation. Our approach, for the first time, presents a generic framework for both language and audio references, and enhances the interaction between temporal frames and multi-modal signals. In detail, we adopt a DETR-like [5] encoder-decoder transformer, which serves as the basic architecture to process visual information within different frames. On top of this, we introduce two attention-based modules respectively for low-level multi-modal temporal aggregation (MTA), and high-level multi-object temporal interaction (MTI). Firstly before the transformer, we utilize the encoded multi-modal references as queries to aggregate informative visual and temporal features via the MTA module. We concatenate the visual features of adjacent frames and adopt sequential attention blocks for multi-modal tokens to progressively capture temporal visual cues of different image scales. This contributes to better low-level cross-modal alignment and temporal consistency. Then, we regard the multi-modal tokens after MTA as object queries and feed them into the transformer for frame-wise decoding. After that, we apply the MTI module to conduct inter-frame object-wise interaction, and maintain a set of video-wise query representation for associating objects across frames inspired by [24]. Such a module enhances the instance-level temporal communication and benefits the visual correspondence for segmenting the same object in a video. Finally, we utilize a segmentation head following previous works [57, 58] to output the final object mask referred by multi-modality input.

To evaluate our effectiveness, we conduct extensive experiments on several popular benchmarks for multi-modal VOS. RVOS with language reference (Ref-YouTube-VOS [50] and Ref-DAVIS 2017 [30]), and one benchmark for AV-VOS with audio reference (AVSBench [71]). On Ref-YouTube-VOS [50] and Ref-DAVIS 2017 [30] with language references, MUTR surpasses the state-of-the-art method ReferFromer [57] by +4.2% and +4.1% $\mathcal{J}\&\mathcal{F}$ scores, respectively. On AV-VOS [71] with audio references, we also outperform Baseline [71] by +4.2% $\mathcal{J}\&\mathcal{F}$ score.

Overall, our contributions are summaried as follows:

- For the first time, we present a unified transformer architecture, termed as MUTR, to tackle video object segmentation referred by multi-modal inputs, i.e., either language or audio.
- To better align the temporal information with multi-modal signals, we propose two attention-based modules, MTA and MTI, respectively for low-level multi-scale aggregation and high-level multi-object interaction, achieving superior cross-modal understanding in a video.
- On benchmarks of two modalities, our approach both achieves state-of-the-art results, e.g., +4.2 % and +4.1% $\mathcal{J}\&\mathcal{F}$ for Ref-YouTube-VOS and Ref-DAVIS 2017, +4.2% $\mathcal{J}\&\mathcal{F}$ for AV-VOS. This fully indicates the significance and generalization ability of MUTR.

2 Related Work

Referring video object segmentation. Referring Video Object Segmentation (R-VOS) introduces the language expression for target object tracking and segmentation, so this will be a more challenging task than semi-supervised video object segmentation, which provides the ground truth for the first frame mask. Existing R-VOS methods can be broadly classified into three categories. One of the most straightforward ideas is to apply referring image segmentation methods [11, 16, 26, 61, 55] independently to video frames, such as RefVOS [3]. It is obvious that this approach disregards the valuable long-temporal information in videos, which makes it difficult to process common video challenges like object disappearance in reproduction and motion occlusion, blurring, etc. Another approach involves propagating the target mask detected from several key frames throughout the video and selecting the object to be segmented based on a visual grounding model [29, 39, 63, 69]. Although this method makes use of the temporal information to some extent and achieves significant performance improvement, its complex multi-stage training approach is not desirable. The very recent

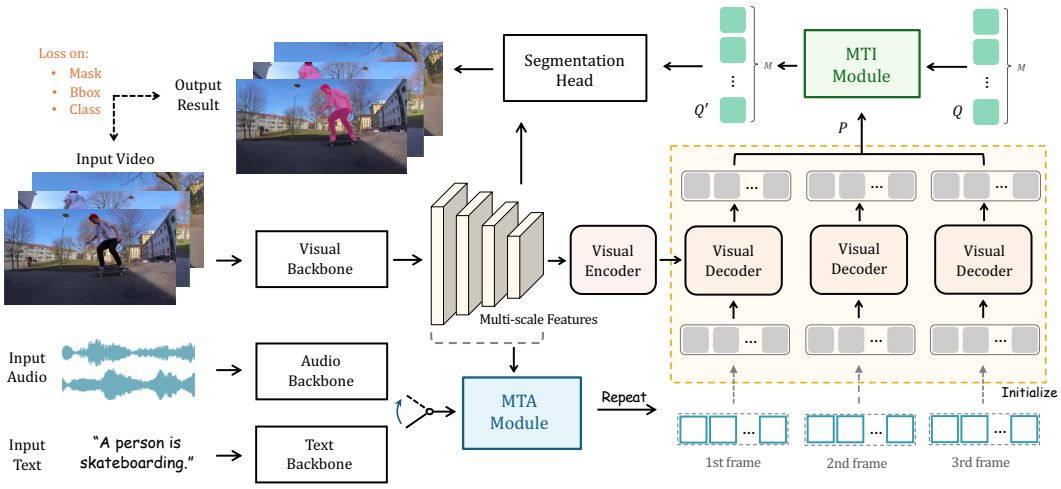

Figure 1: **The Overall Pipeline of MUTR for referring video object segmentation.** We present a unified transformer architecture to tackle video object segmentation referred by multi-modal inputs. We propose MTA module and MTI module for low-level multi-scale aggregation and high-level multi-object interaction, respectively.

work MTTR [4] and ReferFormer [57] have employed query-based mechanisms. Nevertheless, they are end-to-end transformer frameworks, they perform R-VOS task utilizing image-level reference segmentation and lose the most valuable temporal information. In contrast to these methods, our unified framework fully explores video-level visual-attended language information for low-level temporal aggregation.

Audio-visual video object segmentation. Audio-visual video object segmentation (AV-VOS) is a typical and challenging problem of predicting pixel-level individual positions based on a given sound signal. There is little previous work on audio-visual video object segmentation. Until recently [71] proposed the audio-visual video object segmentation dataset and a mechanism to this task. It adopts cross attention to learn the audio-visual correspondence. Different from it, [45] is based on the recent visual foundation model Segment Anything Model [31, 67] to achieve audio-visual segmentation. However, all of them lack the temporal alignment between multi-modal information.

Transformer. Transformer [53] was first proposed as a sequence-to-sequence attention-based building block for machine translation. Due to its powerful capabilities in global context modeling, it has become the cornerstone for most natural language process (NLP) [1, 2, 66, 17] and computer vision (CV) [14, 10, 22] domains. More recently, DETR [5] and its variants [72, 70, 19] introduced the query-based paradigm for simplifying the traditional object detection pipeline to end-to-end while achieving performance comparable to that of CNN-based detectors [21]. DETR reformulates detection as a set prediction task and employs a set of learnable object queries as candidates to predict bounding boxes. MonoDETR [68] introduces a depth-guided DETR for monocular 3D object detection, and iQuery [8] utilizes DETR-based architecture for audio-visual video segmentation. VisTR [54] extends the idea behind DETR to the domain of video instance segmentation (VIS) , which solves the problem by supervising all instances of each frame in a video at the sequence level decoding manner. Inspired by these works, our work is also based on the query-based paradigm, but considers the alignment of multi-modal information on temporal, which is implemented by capturing informative visual cues from consecutive video frames, and then regard them as queries for the transformer.

3 Method

In this section, we illustrate the details of our MUTR for multi-modal video object segmentation. We first describe the overall pipeline in Section 3.1. Then, in Section 3.2 and Section 3.3, we respectively elaborate on the proposed designs of the multi-scale temporal aggregation module (MTA), and multi-object temporal interaction module (MTI).

3.1 Overall Pipeline

The overall pipeline of MUTR is shown in Figure 1. We adopt a DETR-based [5] transformer as our basic architecture, including a visual backbone, a visual encoder and decoder, on top of which, two modules MTA and MTI are proposed for temporal multi-modal interaction. In this section, we successively introduce the pipeline of MUTR for video object segmentation.

Feature Backbone. Given an input video-text/audio pair, we first sample T frames from the video clip, and utilize the visual backbone and a pre-trained text/audio backbone to extract the image and multi-modal features. Specifically, we utilize ResNet [23] or Swin Transformer [36] as the visual backbone, and obtain the multi-scale visual features of the $2^{nd}, 3^{rd}, 4^{th}$ stages. Concurrently, for the text reference, we employ an off-the-shelf language model, RoBERTa [35], to encode the linguistic embedding tokens. For the audio reference, we first process it as a spectrogram transform via a short-time Fourier Transform and then feed it into a pre-trained VGGish [25] model. After the text/audio encoding, a linear projection layer is adopted to align the multi-modal feature dimension with the visual features. Note that, following previous work [57], we adopt an early fusion module in the visual backbone to inject preliminary text/audio knowledge into visual features.

MTA Module. On top of feature extraction, we feed the visual and text/audio features into the multi-scale temporal aggregation module (MTA). We concatenate the visual features of adjacent frames, and adopt cascaded cross-attention blocks to enhance the multi-scale and multi-modal feature fusion, which is specifically described in Section 3.2.

Visual Encoder-decoder Transformer. The basic transformer consists of a visual encoder and a visual decoder, which processes the video in a frame-independent manner to focus on the feature fusion within a single frame. In detail, the visual encoder adopts vanilla self-attention blocks to encode the multi-scale visual features. The visual decoder regards the encoded visual features as the key and value, and the output references from the MTA module as learnable object queries for decoding. Unlike the randomly initialized queries in traditional DETR [5], ours are input-conditioned ones obtained via MTA module, which contains video-level multi-modal prior knowledge. With the visual decoder, the object queries gain rich instance information, which provide effective cues for the final segmentation process.

MTI Module. After the visual transformer, a multi-object temporal interaction (MTI) module is proposed for object-wise interaction, which is described in Section 3.3. In detail, we utilize an MTI encoder to communicate temporal features of the same object in different views. Then an MTI decoder is proposed to grasp information into a set of video-wise query representations for associating objects across frames, inspired by [24].

Segmentation Head and Loss Function. On top of the components introduced above, we obtain the final mask predictions from the extracted multi-modal features via a segmentation head. We follow previous works [57, 58] to design the segmentation head that contains a bounding box head, a classification head, and a mask head. Then, we find the best assignment from the predictions of MUTR by using Hungarian Matching [5]. During training, we calculate three losses in MUTR, which are focal loss [34] $\mathcal{L}_{cls}$ on the predictions of referred object sequence, $\mathcal{L}_{\text{box}}$ on the bounding box of predicted instance, and $\mathcal{L}_{\text{mask}}$ on the predicted object masks. In detail, $\mathcal{L}_{\text{box}}$ is the combination of L_1 loss and GIoU loss [49], and $\mathcal{L}_{\text{mask}}$ is the summation of the Dice [44] and binary focal loss. The whole loss function is formulated as

$$\mathcal{L} = \lambda_{cls}\, \mathcal{L}_{cls} + \lambda_{box}\, \mathcal{L}_{box} + \lambda_{mask}\, \mathcal{L}_{mask}\ , \tag{1}$$

where λ_{cls}, λ_{box} and λ_{mask} denote the weights for $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$ and $\mathcal{L}_{mask}$. See Section 4.2 for detailed settings.

3.2 Multi-scale Temporal Aggregation

To boost both the multi-modal and multi-frame feature fusion, we introduce **M**ulti-scale **T**emporal **A**ggregation module for low-level temporal aggregation. The proposed MTA module generates a set of object queries that contain multi-modal knowledge for subsequent transformer decoding.

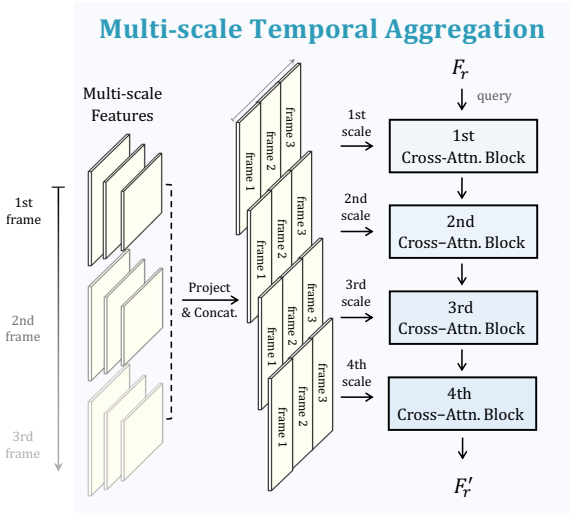

Figure 2: **Multi-scale Temporal Aggregation.** low-level multi-modal temporal aggregation, we propose MTA module for inter-frame interaction, which generates tokens with multi-modal knowledge as the input queries for transformer decoding.

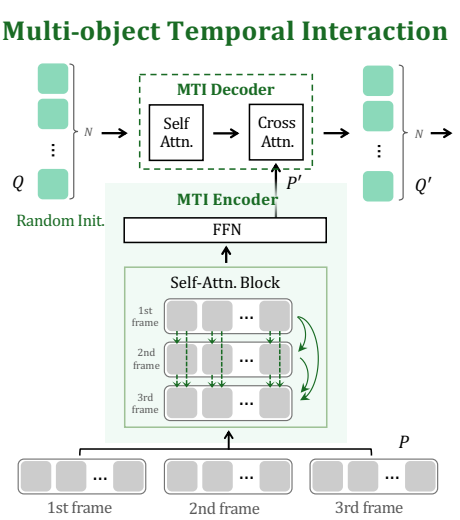

Figure 3: **Multi-object Temporal Interaction.** We introduce MTI module for inter-frame object-wise interaction, and maintain a set of video-wise query representations for associating objects across frames.

Multi-scale Temporal Transform. As shown in Figure 2, the MTA module take the text/audio features F_r, and multi-scale visual features as input, i.e., the extracted features of $2^{nd}, 3^{rd}, 4^{th}$ stages from the visual backbone. We first utilize linear projection layers on the multi-scale features to transform them into the same dimension. Specifically, we separately utilize 1×1 convolution layers on the $2^{nd}, 3^{rd}, 4^{th}$ scale features, and an additional 3×3 convolution layer on the 4^{th} stage features to obtain the 5^{th} scale features. We denote the projected features as $\{F_{vj}^i\}$, where $2 \leq i \leq 5,\ 1 \leq j \leq T$ represent the stage number and frame number. After that, we concatenate the visual features of adjacent frames for each scale, formulated as

$$F_v^i = \text{Concat}(F_{v1}^i,\ F_{v2}^i,\ ...,\ F_{vj}^i,\ ...,\ F_{vT}^i), \tag{2}$$

where $2 \leq i \leq 5,\ 1 \leq j \leq T$, F_{vj}^i represents the projected j^{th} frame features of i^{th} scale, and $\{F_v^i\}_{i=2}^5$ is the final transformed multi-scale visual feature. Then, the resulted multi-modal temporal features are regarded as the key and value in the following cross-attention blocks.

Multi-modal Cross-attention. On top of this, we adopt sequential cross-attention mechanisms for multi-modal tokens to progressively capture temporal visual cues of different image scales. We adopt four cross-attention blocks that are assigned to each scale respectively for multi-scale temporal feature extracting. In each attention block, the text/audio features serve as the query, while the multi-scale visual features serve as the key and value. We formulate it as

$$F_f = \text{Block}_{i-1}(F_r,\ F_v^i,\ F_v^i),\ 2 \leq i \leq 5, \tag{3}$$

where Block represents the sequential cross-attention blocks in MTA module, F_f is the output multi-modal tokens that contain the multi-modal information.

After that, we simply repeat the class token of F_f for $T \times N$ times, where T is the frame number and N is the query number. We adopt them as the initialized queries fed into the visual transformer for frame-wise decoding. With the proposed MTA module, the pre-initialized input queries obtain prior multi-scale knowledge and temporal information for better multi-modal alignment during subsequent decoding.

3.3 Multi-object Temporal Interaction

Since the visual transformer adopts a frame-independent manner and fails to interact information among multiple frames, we further introduce a **M**ulti-object **T**emporal **I**nteraction module to conduct

inter-frame object-wise interaction. This module enhances the high-level temporal communication of objects, and benefits the visual correspondence for effective segmentation. The details of MTI module are shown in Figure 3, which consists of an MTI encoder and an MTI decoder.

MTI Encoder. We obtain the object query outputs P of each frame from the transformer decoder, and feed them into the MTI encoder, which contains a self-attention layer to conduct object-wise interaction across multiple frames, and a feed-forward network layer for feature transformation. To achieve more efficient implementation, we adopt shifted window-attention [36] with linear computational complexity in the self-attention layer. The process of MTI encoder is formulated as

$$P' = \mathrm{MTI\_Encoder}(P) \tag{4}$$

where $\mathrm{MTI\_Encoder}$ denotes the MTI encoder, and P' is the outputs of MTI encoder.

MTI Decoder. Based on the MTI encoder, we maintain a set of video-wise query Q for associating objects across frames, which are randomly initialized. We regard the outputs from MTI encoder as the key and value, and feed them and video-wise queries Q into MTI decoder for video-wise decoding. The MTI decoder consists of a self-attention layer, a cross-attention layer, and a feed-forward network layer. We it them as

$$Q' = \mathrm{MTI\_Decoder}(Q,\ P',\ P') \tag{5}$$

where $\mathrm{MTI\_Decoder}$ represents the MTI decoder, Q' is the outputs of MTI decoder. In this way, the proposed MTI module promotes high-level temporal fusion and enhances the connection and interaction of the same objects in different frames, which further contributes to effective segmentation.

4 Experiments

In Section 4.1, we first introduce the evaluation datasets and metrics. Then, we describe the implementation details in Section 4.2. After that, we present our quantitative results and qualitative results on R-VOS and AV-VOS benchmarks in Section 4.3 and 4.4. Finally, in Section 4.5 we conduct extensive ablation studies on Ref-YouTube-VOS [50] dataset.

4.1 Datasets and Metrics.

Datasets. We evaluate the effectiveness of MUTR on two common-used R-VOS benchmarks, i.e., Ref-DAVIS 2017 [30] and Ref-YouTube-VOS [50], and one challenging Audio-Visual Segmentation (AVS) benchmark AVSBench [71]. Ref-DAVIS 2017 is generated from DAVIS [47] by providing more than $1.5k$ referring descriptions with 90 videos in total, which are divided into two subsets: training set with 60 videos and val set with 30 videos. Ref-YouTube-VOS is a large-scale benchmark that contains 3,978 videos and $15k$ referring descriptions with 3,471/202/305 videos in train/validation/test sets. The test set is only available during the competition, so all our experimental evaluation results are performed on the validation set. AVSBench includes 4,932 videos, which are divided into 3,452/740/740 videos for train/val/test subsets.

Evaluation Metrics. We adopt the standard evaluation protocol [46, 47] such as the region accuracy $\mathcal{J}$, boundary accuracy $\mathcal{F}$ and their average value ($\mathcal{J}\&\mathcal{F}$) for all benchmarks. Ref-DAVIS 2017[1] and AVSBench[2] are evaluated by their official evaluation code, respectively. Ref-YouTube-VOS is evaluated by submitting the results to the official evaluation server[3].

4.2 Implementation details.

Following prior works [57], we evaluate our models under various backbones including: ResNet [23], Swin Transformer [36] and Video Swin Transformer [37] and multi-modality inputs such as text or audio to verify the effectiveness of our method. In the visual encoder-decoder transformer, we adopt 4 encoder layers and 4 decoder layers, which are the same as Referformer [57]. The MTA module is composed of a 3-layer encoder and 3-layer decoder. We set the number of multi-modal queries to 5.

[1] https://github.com/davisvideochallenge/davis2017-evaluation
[2] https://github.com/OpenNLPLab/AVSBench
[3] https://codalab.lisn.upsaclay.fr/competitions/3282

Table 1: **Performance of MUTR on Ref-YouTube-VOS and Ref-DAVIS 2017 Datasets.** We report the results of MUTR and prior works on multiple backbones, where our MUTR shows the *state-of-the-art* performance on all datasets.

Method	Backbone	Ref-YouTube-VOS			Ref-DAVIS 2017		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [62]	ResNet-50	34.9	33.3	36.5	34.7	32.2	37.2
CMSA + RNN [62]		36.4	34.8	38.1	40.2	36.9	43.5
URVOS [50]		47.2	45.3	49.2	51.5	47.3	56.0
LBDT-4 [13]		48.2	50.6	49.4	-	-	-
YOFO [32]		48.6	47.5	49.7	53.3	48.8	57.9
MRLR [56]		48.4	51.0	49.7	57.9	53.9	62.0
ReferFormer [57]		58.7	57.4	60.1	61.1	58.0	64.1
MUTR		**61.9**	**60.4**	**63.4**	**65.3**	**62.4**	**68.2**
CITD [33]	ResNet-101	56.4	54.8	58.1	-	-	-
ReferFormer [57]		59.3	58.1	60.4	61.0	58.1	63.8
MUTR		**63.6**	**61.8**	**65.4**	**65.3**	**61.9**	**68.6**
ReferFormer [57]	Swin-L	64.2	62.3	66.2	63.9	60.8	67.0
MUTR		**68.4**	**66.4**	**70.4**	**68.0**	**64.8**	**71.3**
MTTR [4]	Video-Swin-T	55.3	54.0	56.6	-	-	-
MANet [9]		55.6	54.8	56.5	-	-	-
ReferFormer [57]		62.6	59.9	63.3	62.8	60.8	67.0
MUTR		**64.0**	**62.2**	**65.8**	**66.5**	**63.0**	**70.0**
ReferFormer [57]	Video-Swin-S	63.3	61.4	65.2	62.3	58.8	65.8
MUTR		**65.1**	**63.0**	**67.1**	**66.1**	**62.5**	**69.8**
VLT [12]	Video-Swin-B	63.8	61.9	65.6	61.6	58.9	64.3
ReferFormer [57]		64.9	62.8	67.0	64.3	60.7	68.0
MUTR		**67.5**	**65.4**	**69.6**	**66.4**	**62.8**	**70.0**

We train the model with AdamW [38] optimizer on 8 A100 GPUs. The initial learning rate of the transformer and backbone are 1×10^{-4} and 6×10^{-6}. The weight decay is set to 5×10^{-4}. We set the batch size to 2 and 8 on R-VOS and AV-VOS, respectively. For losses, the coefficients are set as $\lambda_{\text{cls}} = 2$, $\lambda_{L_1} = 5$, $\lambda_{\text{giou}} = 2$, $\lambda_{\text{dice}} = 5$, $\lambda_{\text{focal}} = 2$. For all datasets, the frames are downsampled that the lower resolution with the shortest side ranges from 288 to 512, and the longest side is smaller than 640, referring to the setting in [57].

R-VOS. All of our models are trained for 12 epochs on the mixed data from Ref-YouTube-VOS and Ref-COCO [64]. The training video clip consists of 5 frames, while for the static images in Ref-COCO, a synthetic video sequence of 5 frames is generated using various data augmentations, including affine and perspective transformations.

The learning rate is scaled by the factor 0.1 on the 8th and 10th epoch.

AV-VOS. Since only the first video frame in the AVSBench dataset is annotated, we perform data augmentations to the first frame to generate a pseudo video sequence of 5 frames for training. The models are trained for a total of 40 epochs, and then the learning rate decreases at the 30th epoch and the 35th epoch, where the drop coefficient is 0.1.

4.3 Quantitative Results

Ref-YouTube-VOS. As show in Table 1, MUTR outperforms the previous state-of-the-art methods by a large margin under on all datasets. On Ref-YouTube-VOS, MUTR with a lightweight backbone ResNet-50 achieves the superior performance with overall $\mathcal{J}\&\mathcal{F}$ of 61.9%, an improvement of +3.2% than the previous state-of-the-art method Referformer. By adopting a more powerful backbone Swin-Transformer [36], MUTR improves the performance to $\mathcal{J}\&\mathcal{F}$ 68.4%, which is +4.2% than the previous method ReferFormer [57]. Using a more strong backbone, our method has a higher

Table 2: **Performance of MUTR on AVSBench Dataset.** MUTR surpasses the *state-of-the-art* method.

Methods	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
LVS [7]	ResNet-18	44.5	37.9	51.0
MSSL [48]	ResNet-18	55.6	44.9	66.3
3DC [40]	ResNet-152	66.5	57.1	75.9
SST [15]	ResNet-50	73.2	66.3	80.1
iGAN [42]	-	69.7	61.6	77.8
LGVT [65]	Swin-B	81.1	74.9	87.3
Baseline [71]	ResNet-50	78.8	72.8	84.8
Baseline [71]	Pvt-v2	83.3	78.7	87.9
MUTR	ResNet-50	83.0	78.6	87.3
	ResNet-101	83.1	78.5	87.6
	Swin-L	85.7	81.5	89.8
	Video-Swin-T	83.0	78.7	87.2
	Video-Swin-S	84.1	79.8	88.3
	Video-Swin-B	85.7	81.6	89.7

Table 3: Ablation Study of the MTA and MTI Modules of MUTR.

MTA	MTI	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
-	-	60.2	58.7	61.7
-	✓	60.0	58.5	61.5
✓	-	61.5	60.1	63.0
✓	✓	**61.9**	**60.4**	**63.4**

Table 4: Ablation Study of Query Number in Visual Transformer and MTI Module.

Query Number	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	61.2	59.6	62.9
3	61.3	59.9	62.7
5	**61.9**	**60.4**	**63.4**
7	61.4	59.9	62.8

percentage of improvement, which better reflects the robustness of our method on the scaled-up model size. To reflect the powerful temporal modeling capability of MUTR, we therefore adopt the video Swin transformer [37] as the backbone, which is a spatial-temporal encoder that can effectively capture the spatial and temporal cues simultaneously, to compensate for the temporal limitations of the ReferFormer as discussed in [27]. It can be observed that our method significantly outperforms the ReferFormer, which demonstrates the effectiveness of the temporal consistency in our model.

Ref-DAVIS 2017. On the Ref-DAVIS 2017, our method also achieves the best results under the same backbone setting. Since ReferFormer [57] does not include the results

on Ref-DAVIS 2017, we report its results using the official pre-trained models provided by ReferFormer.

AV-VOS. Table 2 shows the performance of our MUTR on the AVSBench dataset. MUTR significantly surpasses all the previous best competitors ($\mathcal{J}\&\mathcal{F}$ **83.0% VS 78.8%**) with the same ResNet-50 backbone. We also achieve a new state-of-the-art performance with Swin-L [36] backbone. By employing a stronger backbone, we observe consistent performance improvement of MUTR, indicating the strong generalization of our approach.

4.4 Qualitative Results

The first two columns of Figure 4 visualize some qualitative results in comparison with ReferFormer [57], which lacks inter-frame interaction in terms of temporal dimension. As demonstrated, along with multiple highly similar objects in the video, ReferFormer [57] is easier to misidentifies them. In contrast, our MUTR is able to associate all the objects in temporal, which can better track and segment all targets accurately.

The last column of Figure 4 visualizes the audio-visual result compared with Baseline [71] on AVSBench dataset. With temporal consistency, MUTR can successfully track and segment challenging situations that are surrounded or occluded by similar instances.

4.5 Ablation Studies

In this section, we perform extensive experiments to analyze the main components and hyper-parameters of MUTR. All the ablation experiments are conducted with the ResNet-50 backbone and evaluate their impact by the Ref-YouTube-VOS performance.

Component Effectiveness Study. Table 3 demonstrates effectiveness of the Multi-scale Temporal Aggregation (MTA) and Multi-object Temporal Interaction (MTI) proposed in our framework. The performance will be seriously degraded from 61.9% to 60.2% by removing MTA and MTI modules.

Figure 4: **Qualitative Results of MUTR.** We visualize the results between ReferFormer [57] and MUTR on R-VOS benchmarks and between Baseline [71] and MUTR on AV-VOS benchmark. Compared with ReferFormer, MUTR performs better on temporal consistency when segmenting multiple similar objects, i.e., fire truck in Ref-YouTube-VOS and box in Ref-DAVIS 2017. Also, compared with the baseline of AV-VOS [71] that denoted as 'Baseline' in this figure, MUTR can handle serve occlusion.

Table 5: Ablation Study of MTA Module.

Components		Block Num.	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Multi-scale	Temporal				
✓	-	1	61.3	59.7	62.7
-	✓	1	60.4	58.9	61.9
✓	✓	1	**61.9**	**60.4**	**63.4**
✓	✓	2	60.7	59.3	62.2
✓	✓	3	60.4	59.1	61.7

Table 6: Ablation Study of MTI Module.

Components		Block Num.	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Encoder	Decoder				
✓	-	3	60.3	58.8	61.9
-	✓	3	61.2	60.0	62.6
✓	✓	3	**61.9**	**60.4**	**63.4**
✓	✓	2	61.1	59.5	62.6
✓	✓	1	60.8	59.3	62.3

Ablation Study on MTA. In Table 5, if either the single-scale temporal aggregation or multi-scale aggregation at the image level are adopted, the performance of MUTR would significantly drop to 60.4% and 61.3%, respectively, which demonstrates the necessity of the MTA module. We also ablate the number of MTA blocks. As seen in Table 5, more MTA blocks cannot bring further performance improvement, since (1) not enough videos for training; (2) the embedding space of visual and reference is only 256-dimensional, which is difficult to optimize so many parameters.

Ablation Study on MTI. According to the results in Table 6, the performance of MUTR is improved by using more MTI blocks. A possible reason is that the larger the MTI blocks, the more sufficient temporal communication between instance-level can be performed. Moreover, using only the encoder or decoder, the performance of MUTR would both decline.

Query Number. We also investigate the influence of different numbers of query on the final performance in Table 4. In our default setting, we set the query number N to 5. As shown in Table 4, even in the case of $N = 1$, MUTR still achieves very competitive results. It is observed that more queries enable the model to select from multiple candidate instances, which make it handle more challenging situations when the object to be referred is surrounded by other similar instances. However, as the number of queries increases (e.g. $N = 7$), there is a clear degradation in performance. One possible reason is the imbalance of number between positive and negative samples, since only one object is referenced.

5 Conclusion

This paper proposes a MUTR, a **M**ulti-modal **U**nified **T**emporal transformer for **R**eferring video object segmentation. A simple yet and effective Multi-scale Temporal Aggregation (MTA) is introduced for multi-modal references to explore low-level multi-scale visual information in video-level. Besides, the high-level Multi-object Temporal Interaction (MTI) is designed for inter-frame feature communication to achieve temporal correspondence between the instance-level across the entire video. Aided by the MTA and MTI, our MUTR achieves new state-of-the-art performance on three R-VOS/AV-VOS benchmarks compared to previous solutions. We hope the MTA and MTI will help ease the future study of multi-modal VOS and related tasks (e.g., referring video object tracking and video instance segmentation). We do not foresee negative social impact from the proposed work.

References

[1] Ahmed, K., Keskar, N.S., Socher, R.: Weighted transformer network for machine translation. arXiv preprint arXiv:1711.02132 (2017)

[2] Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16. pp. 319–334. Springer (2021)

[3] Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J., Giro-i Nieto, X.: Refvos: A closer look at referring expressions for video object segmentation. arXiv preprint arXiv:2010.00263 (2020)

[4] Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4985–4995 (2022)

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020)

[6] Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7454–7463 (2019)

[7] Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021)

[8] Chen, J., Zhang, R., Lian, D., Yang, J., Zeng, Z., Shi, J.: iquery: Instruments as queries for audio-visual sound separation. CVPR 2023 (2022)

[9] Chen, W., Hong, D., Qi, Y., Han, Z., Wang, S., Qing, L., Huang, Q., Li, G.: Multi-attention network for compressed video referring object segmentation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4416–4425 (2022)

[10] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)

[11] Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330 (2021)

[12] Ding, H., Liu, C., Wang, S., Jiang, X.: Vlt: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[13] Ding, Z., Hui, T., Huang, J., Wei, X., Han, J., Liu, S.: Language-bridged spatial-temporal interaction for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4964–4973 (2022)

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[15] Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5912–5921 (2021)

[16] Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15506–15515 (2021)

[17] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)

[18] Gao, P., Ma, T., Li, H., Dai, J., Qiao, Y.: Convmae: Masked convolution meets masked autoencoders. arXiv preprint arXiv:2205.03892 (2022)

[19] Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of detr with spatially modulated co-attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3621–3630 (2021)

[20] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)

[21] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)

[22] Guo, Z., Tang, Y., Zhang, R., Wang, D., Wang, Z., Zhao, B., Li, X.: Viewrefer: Grasp the multi-view knowledge for 3d visual grounding with gpt and prototype guidance. arXiv preprint arXiv:2303.16894 (2023)

[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[24] Heo, M., Hwang, S., Oh, S.W., Lee, J.Y., Kim, S.J.: Vita: Video instance segmentation via object token association. arXiv preprint arXiv:2206.04403 (2022)

[25] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017)

[26] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 108–124. Springer (2016)

[27] Hu, Z., Chen, B., Gao, Y., Ji, Z., Bai, J.: 1st place solution for youtubevos challenge 2022: Referring video object segmentation. arXiv preprint arXiv:2212.14679 (2022)

[28] Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4424–4433 (2020)

[29] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)

[30] Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14. pp. 123–141. Springer (2019)

[31] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

[32] Li, D., Li, R., Wang, L., Wang, Y., Qi, J., Zhang, L., Liu, T., Xu, Q., Lu, H.: You only infer once: Cross-modal meta-transfer for referring video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1297–1305 (2022)

[33] Liang, C., Wu, Y., Zhou, T., Wang, W., Yang, Z., Wei, Y., Yang, Y.: Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. arXiv preprint arXiv:2106.01061 (2021)

[34] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

[35] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[36] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

[37] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)

[38] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[39] Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 10034–10043 (2020)

[40] Mahadevan, S., Athar, A., Ošep, A., Hennen, S., Leal-Taixé, L., Leibe, B.: Making a case for 3d convolutions for object segmentation in videos. arXiv preprint arXiv:2008.11516 (2020)

[41] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)

[42] Mao, Y., Zhang, J., Wan, Z., Dai, Y., Li, A., Lv, Y., Tian, X., Fan, D.P., Barnes, N.: Transformer transforms salient object detection and camouflaged object detection. arXiv preprint arXiv:2104.10127 (2021)

[43] Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 630–645 (2018)

[44] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)

[45] Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)

[46] 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. pp. 724–732 (2016)

[47] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)

[48] Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 292–308. Springer (2020)

[49] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)

[50] Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 208–223. Springer (2020)

[51] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)

[52] Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 38–54 (2018)

[53] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[54] Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8741–8750 (2021)

[55] Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)

[56] Wu, D., Dong, X., Shao, L., Shen, J.: Multi-level representation learning with semantic alignment for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4996–5005 (2022)

[57] Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)

[58] Wu, J., Jiang, Y., Bai, S., Zhang, W., Bai, X.: Seqformer: Sequential transformer for video instance segmentation. arXiv preprint arXiv:2112.08275 (2021)

[59] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)

[60] Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. arXiv preprint arXiv:2303.06674 (2023)

[61] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)

[62] Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10502–10511 (2019)

[63] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1307–1315 (2018)

[64] Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)

[65] Zhang, J., Xie, J., Barnes, N., Li, P.: Learning generative vision transformer with energy-based latent space for saliency prediction. Advances in Neural Information Processing Systems **34**, 15448–15463 (2021)

[66] Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)

[67] Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)

[68] Zhang, R., Qiu, H., Wang, T., Xu, X., Guo, Z., Qiao, Y., Gao, P., Li, H.: Monodetr: Depth-guided transformer for monocular 3d object detection. arXiv preprint arXiv:2203.13310 (2022)

[69] Zhang, Y., Yuan, L., Guo, Y., He, Z., Huang, I.A., Lee, H.: Discriminative bimodal networks for visual localization and detection with natural language queries. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 557–566 (2017)

[70] Zheng, M., Gao, P., Zhang, R., Wang, X., Li, H., Dong, H.: End-to-end object detection with adaptive clustering transformer. BMVC 2021 Oral (2020)

[71] Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 386–403. Springer (2022)

[72] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

A Overview

- Appendix B: Additional implementation details.
- Appendix C: Additional experiments.
- Appendix D: Additional visualizations.

B Additional Experiment Details

B.1 Dataset Details

Ref-YouTube-VOS is built upon YouTube-VOS [59] by providing 12,193 language descriptions for the training set of 3,471 videos and 202/305 videos with 2,096 expressions in validation/test set. Note that, the test set targets on competition that the server is currently inaccessible. Each object is annotated with two kinds of referring expressions for the *first-frame* and *full-video*.

Ref-DAVIS 2017 is expanded from DAVIS [47] by providing 1,544 expression sentences describing 205 objects in total. The referred instance is annotated with two annotators and each of them gives the *first-frame* and *full-video* textual description with the same as Ref-YouTube-VOS. For fair comparisons, we report the results by averaging the scores with the same setting referring to ReferFormer [57].

AVSBench is the first pixel-level audio-visual segmentation benchmark that contains 4,932 videos (5 frames each, 10,852 annotated frames) covering 23 categories including instruments, humans, animals, etc. Only the first frame is annotated in the training set, all the test set and validation set are annotated.

B.2 Implementation Details

R-VOS. For the modality of language expression with the number words of L, we employ an off-the-shelf linguistic model, RoBERTa [35], to extract the text feature $F_r \in \mathbb{R}^{L\times C_l}$, where C_l = 768. The learning rate in the training phase is 1×10^{-5}. Note that, for Ref-DAVIS 2017, we directly report the results using the model trained on Ref-Youtube-VOS without fine-tuning referring to [57].

AV-VOS. Given an input of audio clip, the audio features $F_r \in \mathbb{R}^{L\times C_a}$ are extracted from VGGish [25] that is pre-trained on AudioSet [20], where C_a = 128 is the feature dimension. It should be noted that the audio encoder is frozen all the time following [71].

Table 7: **Performance of MUTR on Ref-YouTube-VOS and Ref-DAVIS 2017 Datasets.** We report the results between MUTR and UNINEXT on multiple backbones, where our MUTR shows the *state-of-the-art* performance on all datasets.

Method	Backbone	Ref-YouTube-VOS			Ref-DAVIS 2017		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
UNINEXT [60]	ResNet-50	61.2	59.3	63.0	63.9	59.6	68.1
MUTR		**61.9**	**60.4**	**63.4**	**65.3**	**62.4**	**68.2**
UNINEXT [60]	ConvNext-Large	66.2	64.0	68.4	66.7	62.3	71.1
MUTR		**66.7**	**64.8**	**68.7**	**69.0**	**65.6**	**72.4**
MUTR	ConvMAE-Base	**66.9**	**64.7**	**69.1**	**69.2**	**65.6**	**72.8**

C Additional Experiments

R-VOS. All of our models are trained on the mixed dataset from image referring segmentation datasets Ref-COCO [64], Ref-COCOg [64], Ref-COCO+ [41] and referring video segmentation dataset Ref-YouTube-VOS. UNINEXT [60] is pre-trained on the large-scale object detection dataset

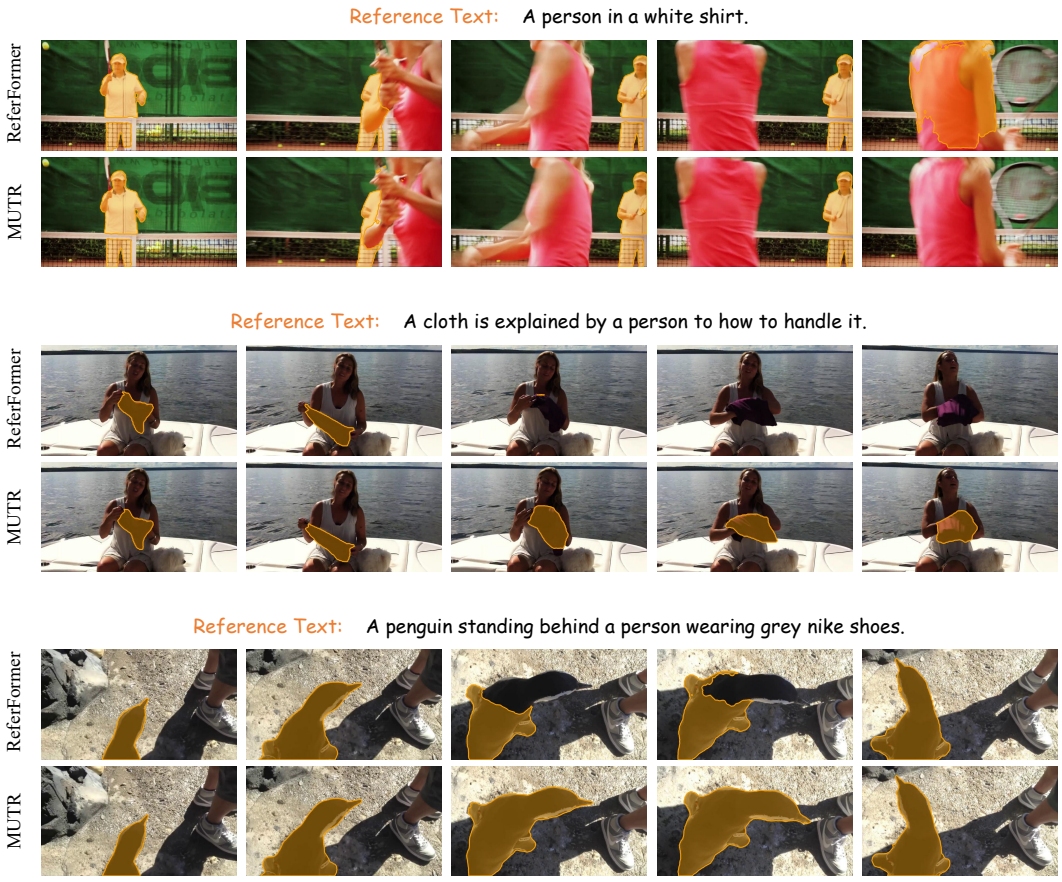

Figure 5: **Qualitative results on Ref-YouTube-VOS between ReferFormer and MUTR.**

Table 8: **Performance of MUTR on AVSBench Dataset.** We report the results between MUTR and Baseline on multiple backbones. $*$ represents the results of our own reproduction.

Methods	Backbone	AVSBench Validation			AVSBench Test		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Baseline [71]	ResNet-50	-	-	-	78.8	72.8	84.8
Baseline* [71]	ResNet-50	76.2	70.9	81.5	77.3	71.9	82.7
MUTR	ResNet-50	82.5	78.3	86.7	83.0	78.6	87.3
	ResNet-101	82.5	78.1	86.8	83.1	78.5	87.6
	Swin-L	85.2	81.2	89.1	85.7	81.5	89.8
	Video-Swin-T	82.9	78.8	83.0	83.0	78.7	87.2
	Video-Swin-S	83.7	79.3	88.0	84.1	79.8	88.3
	Video-Swin-B	85.0	81.1	88.9	85.7	81.6	89.7
	ConvMAE-B	85.5	81.2	89.7	86.3	82.2	90.3

Objects365 [51], and then finetune it on RefCOCO/g/+ and Ref-YouTube-VOS. For comparison, we further evaluate our model under ConvNext-Large and a ViT-based [14] backbone ConvMAE [18]. The results are shown in Table 7.

AV-VOS. The AVSBench benchmark is split into three subsets, 3,452/740/740 for train/val/test sets, respectively. For verifying the effectiveness and robustness of our models, we still evaluate the performance on val set. The performance is shown in Table 8. Note that, the Baseline [71] method does not publicly provide official weights, so we reproduced it with reference to the original paper settings.

Figure 6: **Qualitative results on AVSBench between Baseline and MUTR.**

D Additional Visualizations

Visualizations on Ref-YouTube-VOS. In Figure 5, we visualize the results compared with ReferFormer [57] on Ref-YouTube-VOS benchmark. From Figure 5, MUTR can successfully track and segment the referred instance even in challenging situations, where they are surrounded by similar instances (the first row), much deformation (the second row) and background interface (the third row).

Visualizations on AVSBench. In Figure 6, we visualize the results compared with Baseline [71] on AVSBench benchmark. With temporal consistency (Multi-object Temporal Interaction module), our model can still successfully segment the race car when a person walks past it. Our model can overcome the interference of close proximity to the referred object and similar color textures (Baby Laughter and Driving Bus), thanks in large part to our Multi-scale Temporal Aggregation module.