DMQA Open Seminar

From Referring to Reasoning: The Evolution of Video Object Segmentation

2026.06.12

안채원

Data Mining and Quality Analytics Lab

고려대학교
KOREA UNIVERSITY

발표자 소개

❖ **안채원 (Chaewon An)**

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석사 과정 (2026.03 ~ Present)

❖ **Research Interest**

- Video Object Segmentation
- Multimodal learning

❖ **Contact**

- chaew0n_an@korea.ac.kr

목차

고려대학교 KOREA UNIVERSITY

Introduction

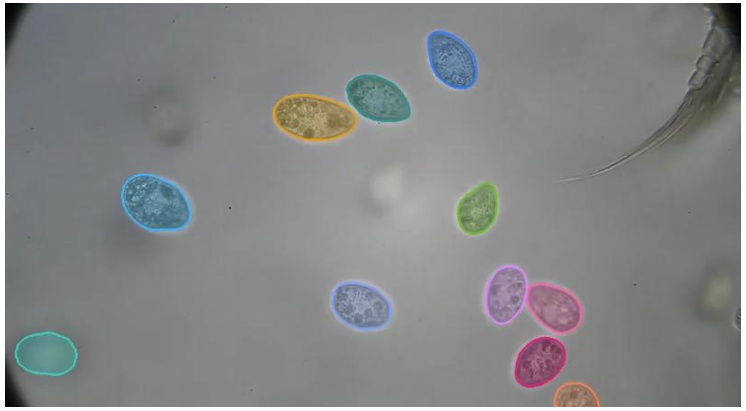
고려대학교
KOREA UNIVERSITY

Introduction

Video Object Segmentation(semi-supervised VOS)

❖ Video Object Segmentation(VOS)이란?

- Video 내 원하는 객체를 프레임 단위로 분할하고 추적하는 기술
- 자율주행, 보안, 의료, 영상 편집 등 다양한 도메인에서 활용됨



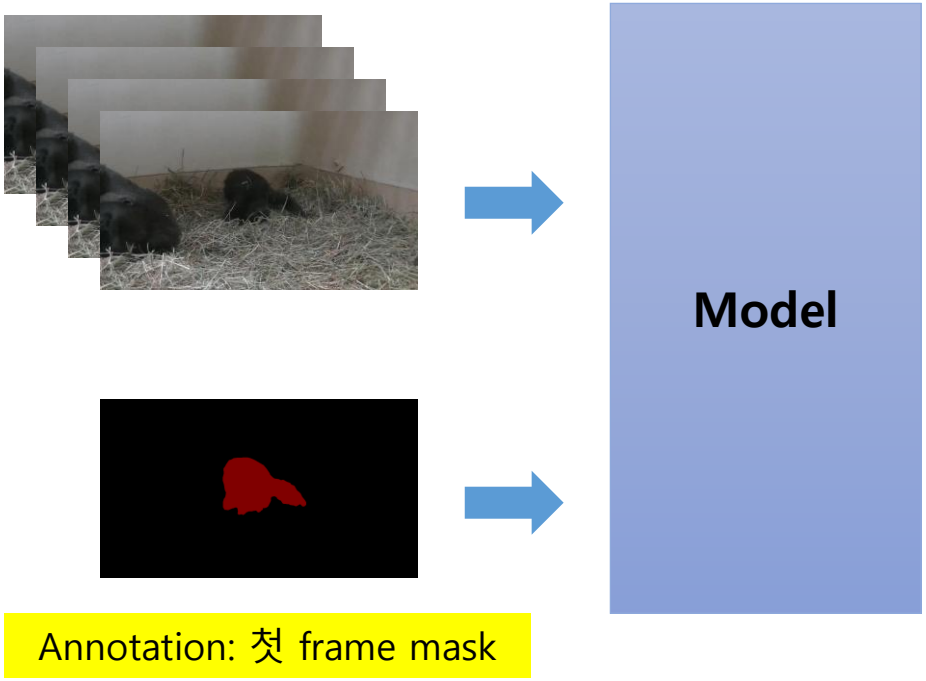
고려대학교 KOREA UNIVERSITY

Introduction

Video Object Segmentation(semi-supervised VOS)

❖ **Video Object Segmentation(VOS)는 어떻게 수행될까?**

- 첫 frame에 대한 mask를 입력 → 연속적인 frame에서 같은 객체를 분리/추적하는 task
- **입력:** Video & Annotation(첫 frame의 추적 대상 mask) → **출력:** 모든 frame에서 segmentation 결과



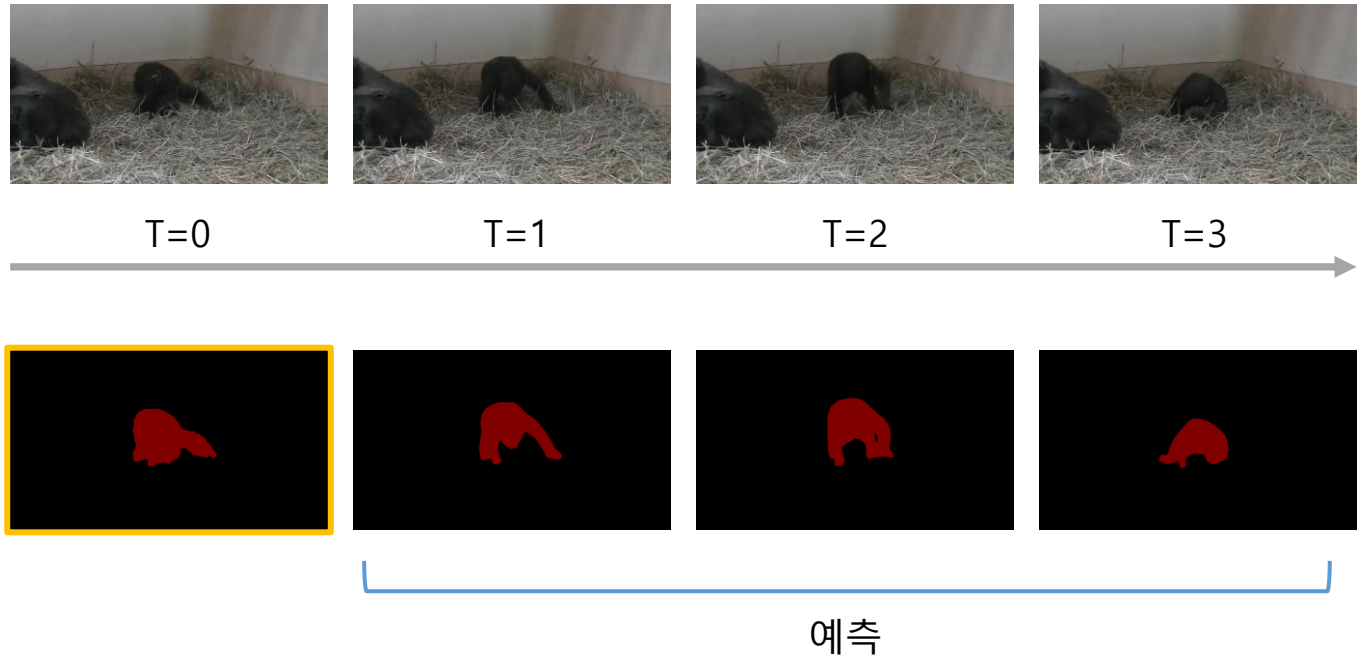
고려대학교 KOREA UNIVERSITY

Introduction

Video Object Segmentation(semi-supervised VOS)

❖ **Video Object Segmentation(VOS)는 어떻게 수행될까?**

- 첫 frame에 대한 mask를 입력 → 연속적인 frame에서 같은 객체를 분리/추적하는 task
- **입력:** Video & Annotation(첫 frame의 추적 대상 mask) → **출력:** 모든 frame에서 segmentation 결과



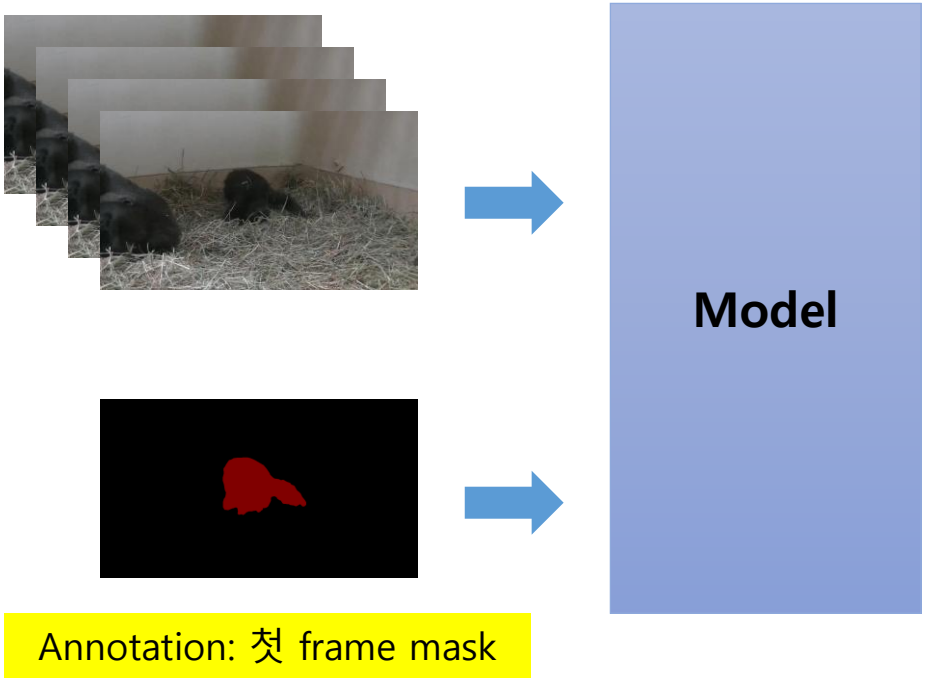
고려대학교 KOREA UNIVERSITY

Introduction

Paradigm shift to language guided VOS

❖ **Annotation을 자연어로 바꿔보자!**

- Annotation은 어떤 객체를 추적할지 정보를 주는 가이드라인 역할
- **Visual guided in VOS →**



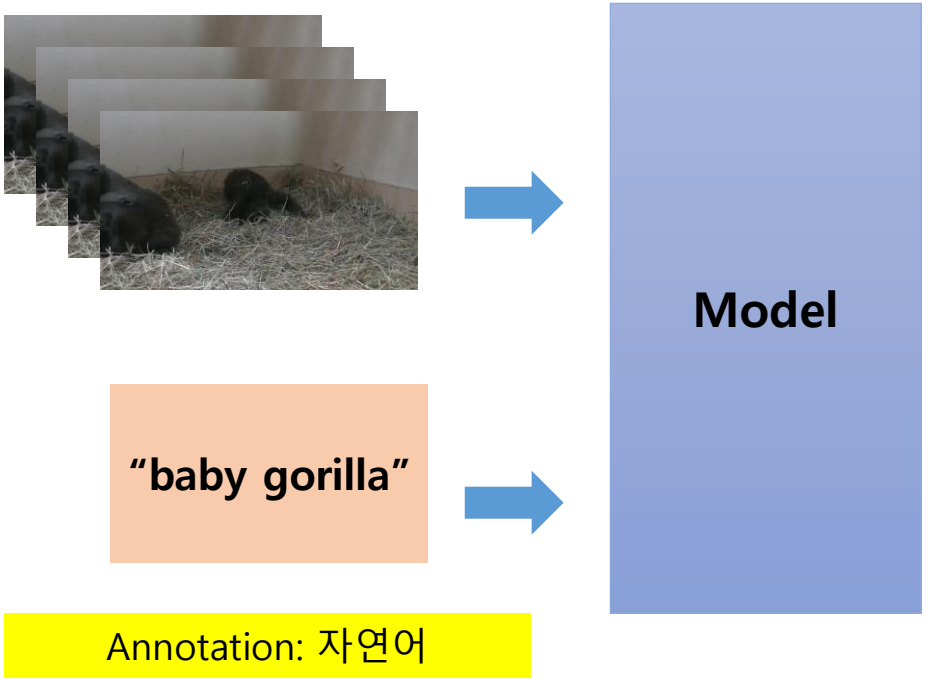
고려대학교 KOREA UNIVERSITY

Introduction

Paradigm shift to language guided VOS

❖ **Annotation을 자연어로 바꿔보자!**

- Annotation은 어떤 객체를 추적할지 정보를 주는 가이드라인 역할
- **Visual guided in VOS → language guided in new VOS task?**



고려대학교 KOREA UNIVERSITY

Introduction

Referring VOS & Reason VOS

❖ **Explicit vs Implicit language guideline**

- **Referring VOS(RVOS):** 명확한 text query로 video의 객체를 분할
- **Reasoning Vos(Reason VOS):** 모호한 text query로 video의 객체를 분할
  - Functional/Affordance Reasoning
  - Causal Reasoning
  - Spatial/Temporal Reasoning
  - World Knowledge Integration

Reason VOS

"the cat on the left"

"the dog jumping"

"the person in red shirt"

Reason VOS

"the object that will protect you from the rain"

"the person second from left"

"the cup that was used most frequently"

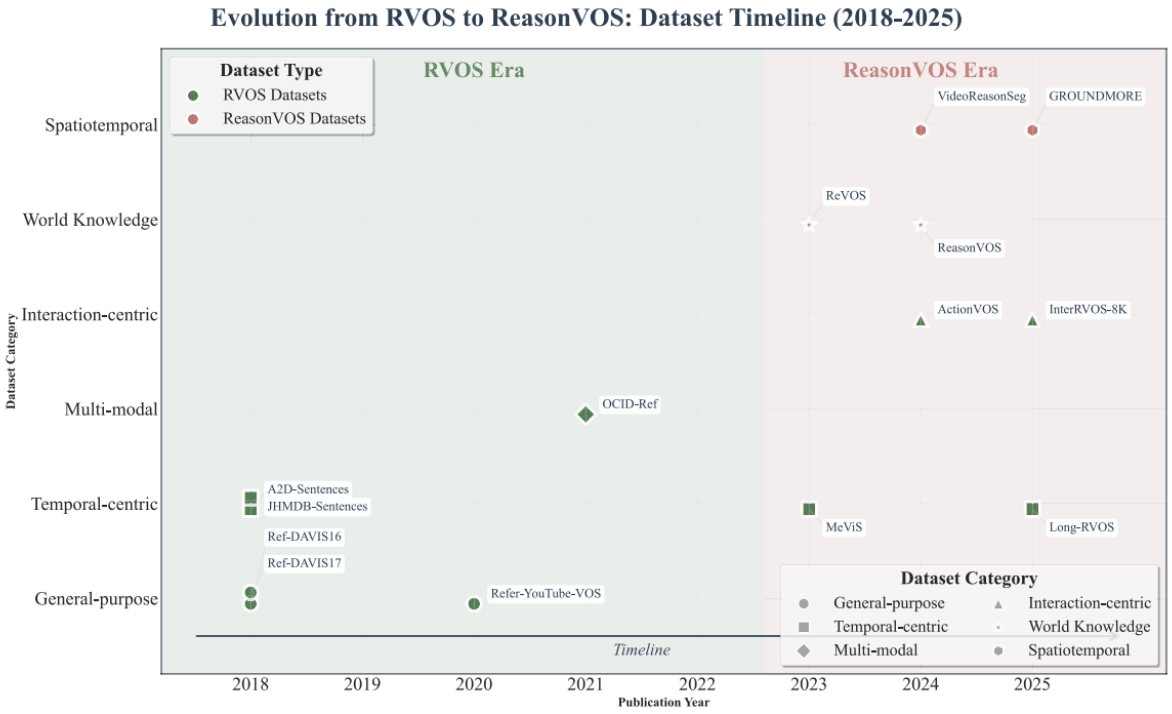
고려대학교 KOREA UNIVERSITY

Introduction

Datasets

❖ **RVOS/Reason VOS task의 등장과 함께 관련 dataset 구축도 활발히 진행 중**

- 2018~2022: 명시적 표현 기반 RVOS 데이터셋 위주
- 2023~2025: 더 복잡해진 RVOS & 추론 기반 Reason VOS 데이터셋



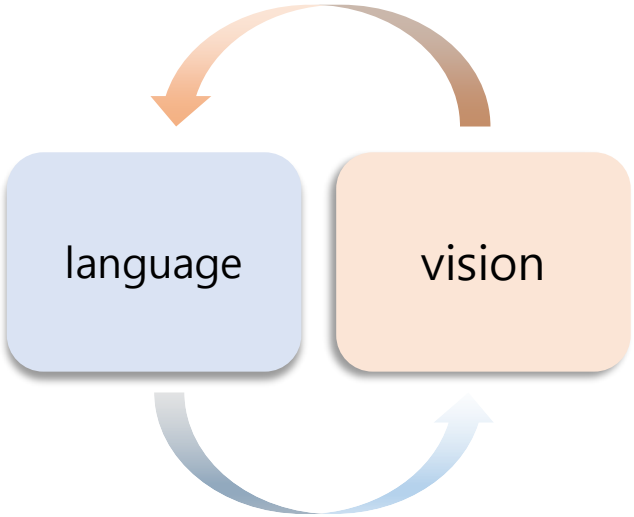
고려대학교 KOREA UNIVERSITY

Introduction

Challenge of RVOS & Reason VOS

❖ **Language guided VOS의 기존 VOS 대비 Challenge는 무엇일까?**

- **RVOS:** cross-modal alignment를 해결 필요
- **Reason VOS:** implicit한 쿼리를 다뤄야 하기 때문에 추가적인 reasoning이 필요



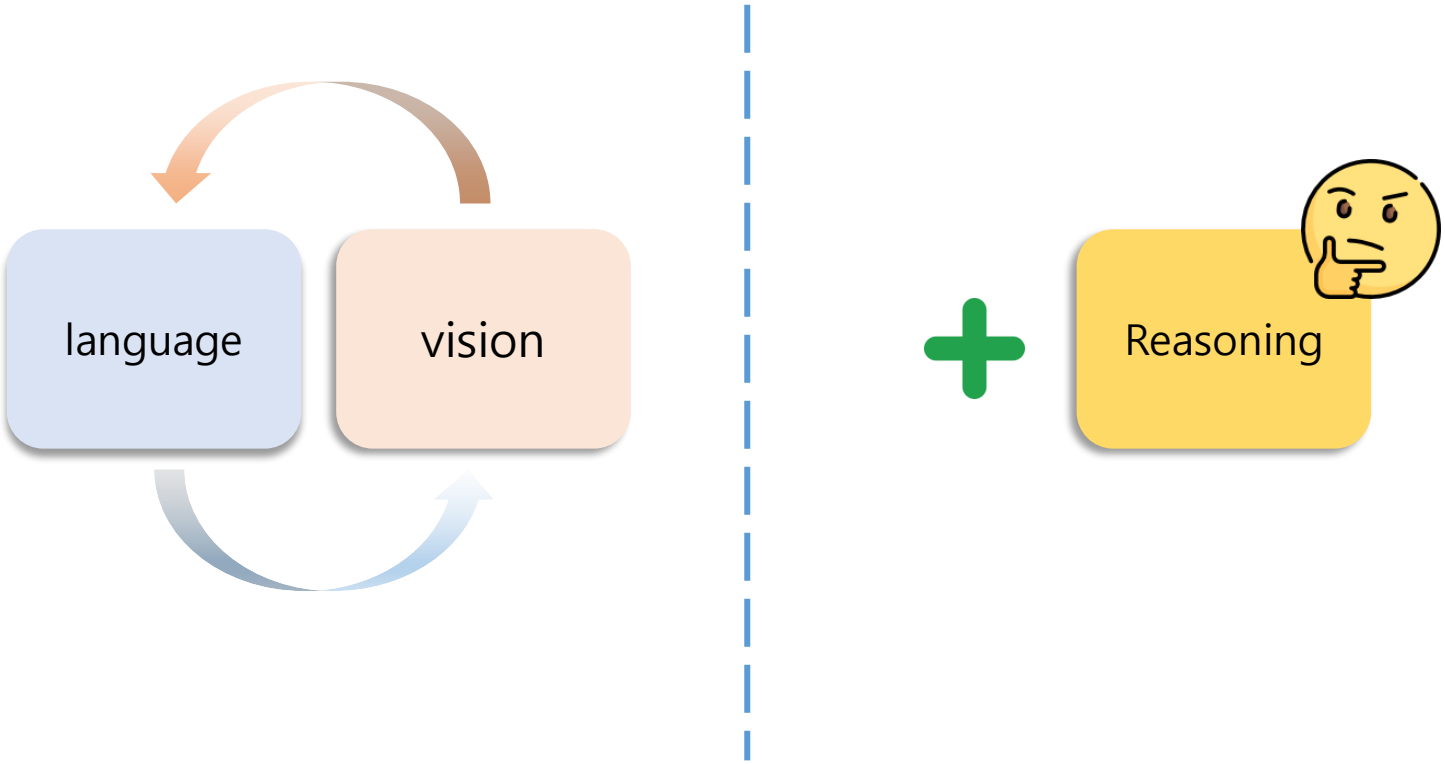
고려대학교 KOREA UNIVERSITY

Introduction

Challenge of RVOS & Reason VOS

❖ **Language guided VOS의 기존 VOS 대비 Challenge는 무엇일까?**

- **RVOS:** cross-modal alignment를 해결 필요
- **Reason VOS:** implicit한 쿼리를 다뤄야 하기 때문에 추가적인 reasoning이 필요



고려대학교 KOREA UNIVERSITY

Algorithms

고려대학교
KOREA UNIVERSITY

Referring VOS

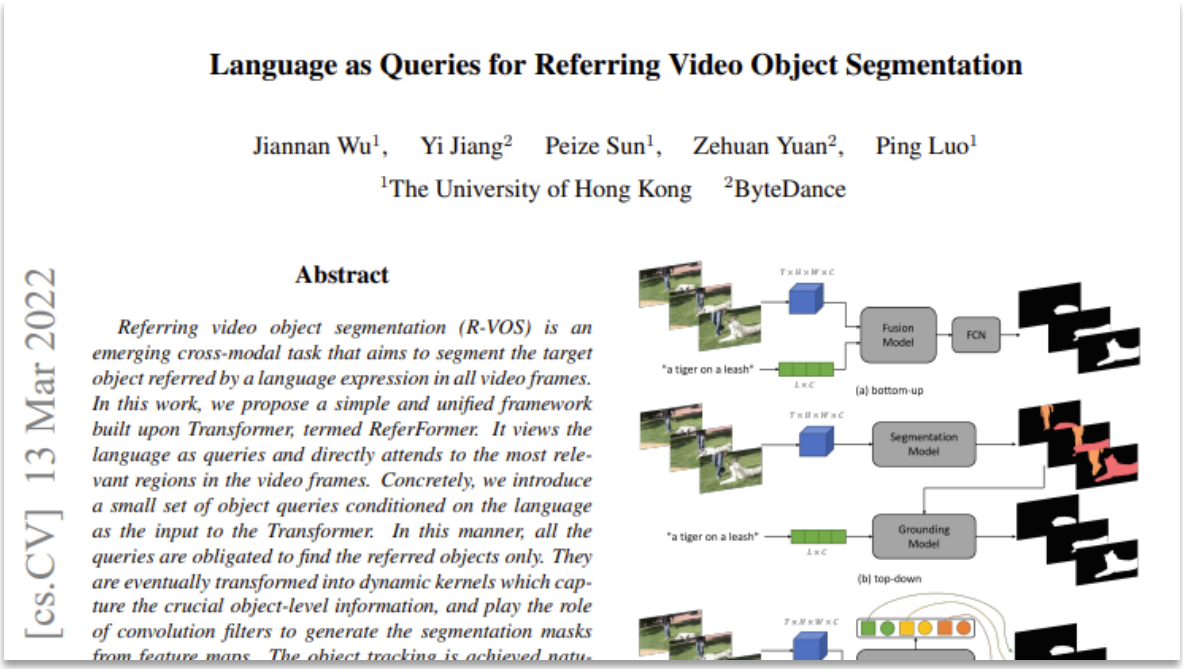
고려대학교
KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Language as Queries for Referring Video Object Segmentation

- 2022 CVPR accept paper
- 334회 인용



Language as Queries for Referring Video Object Segmentation

Jiannan Wu[1], Yi Jiang[2] Peize Sun[1], Zehuan Yuan[2], Ping Luo[1]

[1]The University of Hong Kong [2]ByteDance

[cs.CV] 13 Mar 2022

Abstract

Referring video object segmentation (R-VOS) is an emerging cross-modal task that aims to segment the target object referred by a language expression in all video frames. In this work, we propose a simple and unified framework built upon Transformer, termed ReferFormer. It views the language as queries and directly attends to the most relevant regions in the video frames. Concretely, we introduce a small set of object queries conditioned on the language as the input to the Transformer. In this manner, all the queries are obligated to find the referred objects only. They are eventually transformed into dynamic kernels which capture the crucial object-level information, and play the role of convolution filters to generate the segmentation masks from feature maps. The object tracking is achieved natu-

Wu, J., Jiang, Y., Sun, P., Yuan, Z., & Luo, P. (2022). Language as queries for referring video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4480–4490.

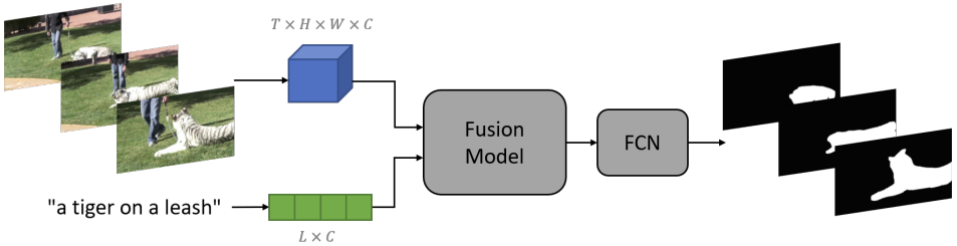
고려대학교 KOREA UNIVERSITY

Referformer

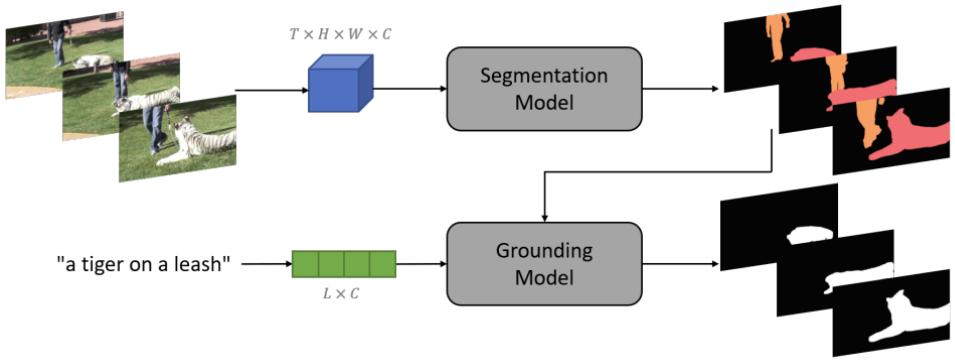
Language as Queries for Referring Video Object Segmentation

❖ Before Referformer

- **Bottom-up:** 시각-언어 정보 fusion → FCN 통과 → 각 픽셀이 지칭된 객체인지 아닌지 분류
- **Top-down:** segmentation 모델로 모든 객체를 검출 → grounding 모델로 자연어 쿼리에 맞는 객체 선택



Bottom-up



Top-down

고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ 이전 방법론의 한계점은 무엇일까?

- **Bottom-up:**
  - 언어를 feature 수준에서 fusion 하는 방식 → 어떤 객체를 찾아야 하는지에 대한 개념 부재
  - 프레임 독립 처리 → 시간적 일관성 부족 → 장면 변화 시 예측 객체가 불일치
- **Top-down:**
  - Segmentation과 grounding이 분리된 구조 → 각 모듈을 따로 학습해야 해서 computational cost가 높음
  - 객체 별 tracklet 생성 후 자연어와 일치하는지 판단 → 언어 정보가 탐색 과정이 아닌 후처리 과정에 도입

고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ **이전 방법론의 한계점은 무엇일까?**

- **Bottom-up:**
  - ➢ 언어를 feature 수준에서 fusion 하는 방식 → 어떤 객체를 찾아야 하는지에 대한 개념 부재

언어 정보를 객체 탐색 과정에 직접 활용할 수 없다!

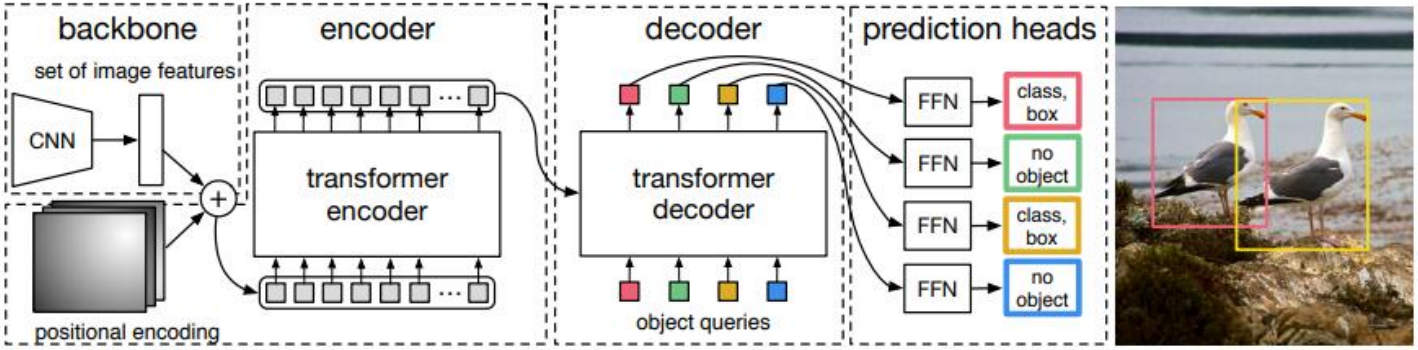
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Referformer는 이전 방법론의 한계를 어떻게 해결했을까?

- **Backgrounds: DETR**은 **object query**를 통해 instance 단위로 객체를 탐색하는 구조
  - ➢ Backbone(CNN)으로 이미지 특징 추출 → Transformer Encoder로 특징 정제
  - ➢ N개의 object query가 **Transformer Decoder**에서 이미지 특징과 attention → 각 query가 객체 하나를 표현하는 instance embedding으로 변환
  - ➢ FFN(Prediction Head)이 각 instance embedding을 받아 class(어떤 객체인지) + bounding box(어디에 있는지) 출력 → instance 단위로 객체 탐지 완료



Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Proceedings of the European Conference on Computer Vision*, 213–229.

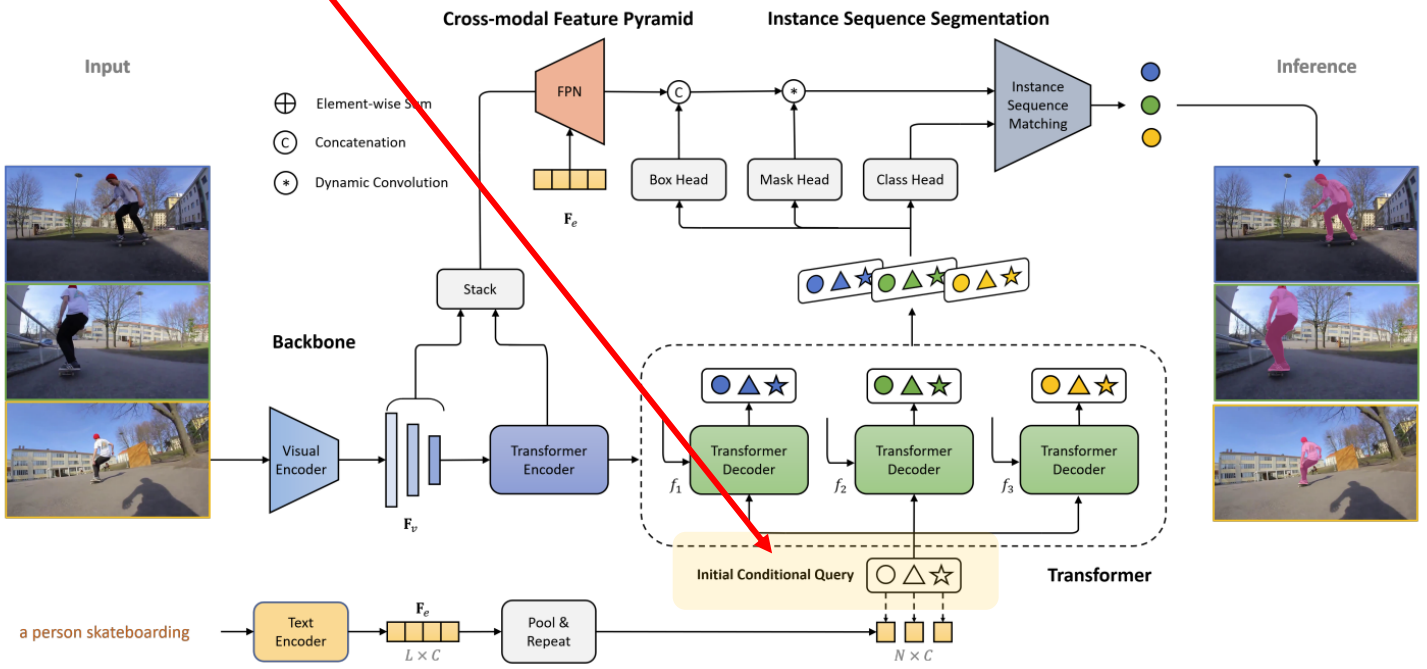
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Referformer는 이전 방법론의 한계를 어떻게 해결했을까?

- Query based mechanism of Transformer(DETR based):
  - ➢ object query = 이미지에서 알아서 객체를 찾는 탐정→ 어떤 객체를 찾을지 임무를 받지 않은 상태



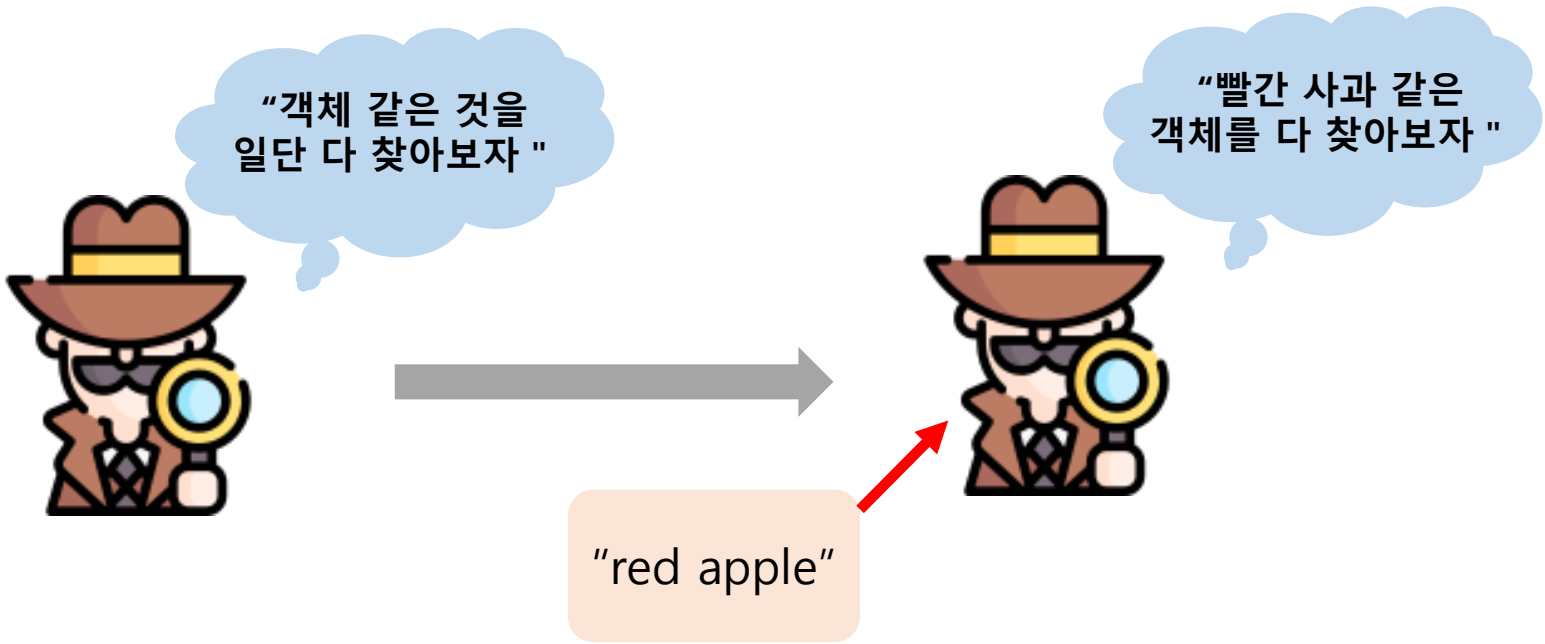
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ **Referformer: language를 query로 활용하자!**

- Query based mechanism of Transformer(DETR based):
  - ➢ **object query** = 이미지에서 알아서 객체를 찾는 탐정→ 어떤 객체를 찾을지 임무를 받지 않은 상태
- **Motivation: 처음부터 언어 단서를 받은 탐정만 투입하면 어떨까?**



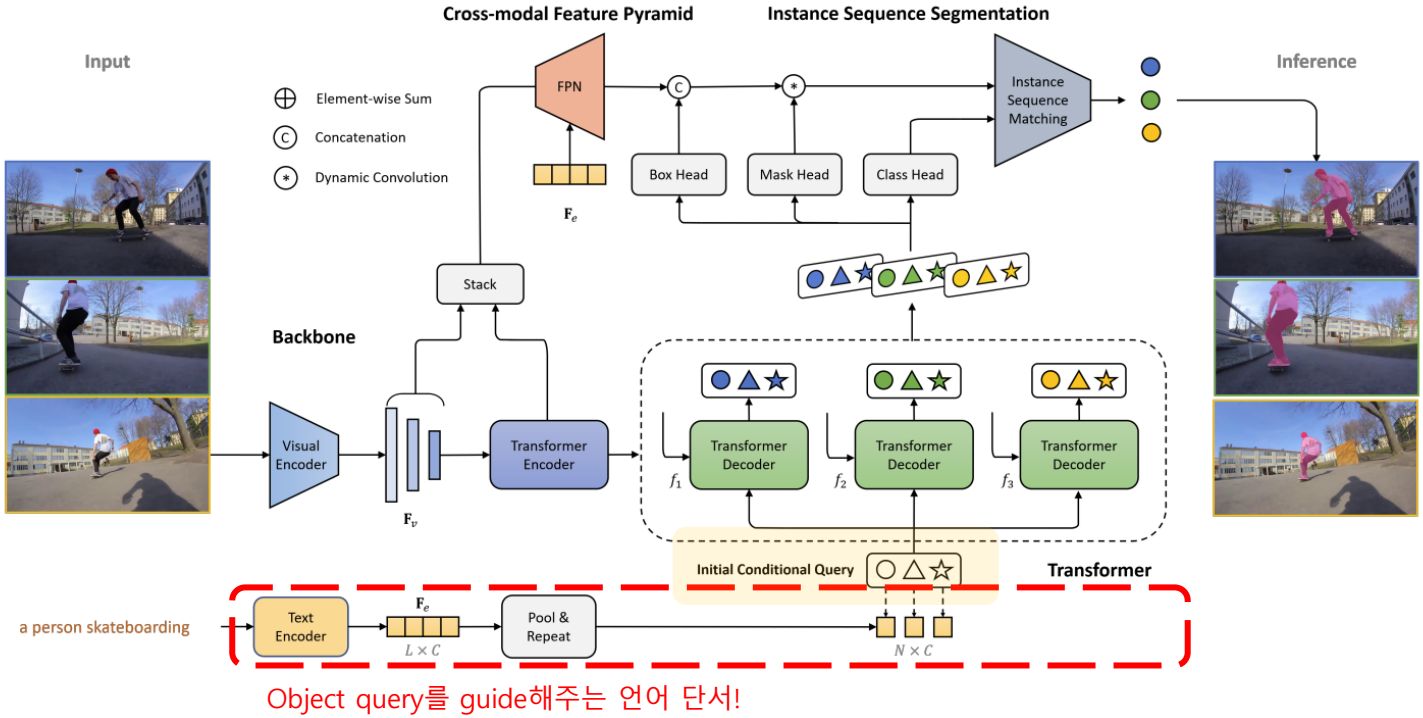
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ **Referformer: language를 query로 활용하자!**

- Query based mechanism of Transformer(DETR based):
  - ➢ **object query** = 이미지에서 알아서 객체를 찾는 탐정→ 어떤 객체를 찾을지 임무를 받지 않은 상태
- **Motivation: 처음부터 언어 단서를 받은 탐정만 투입하면 어떨까?**



Object query를 guide해주는 언어 단서!

고려대학교 KOREA UNIVERSITY

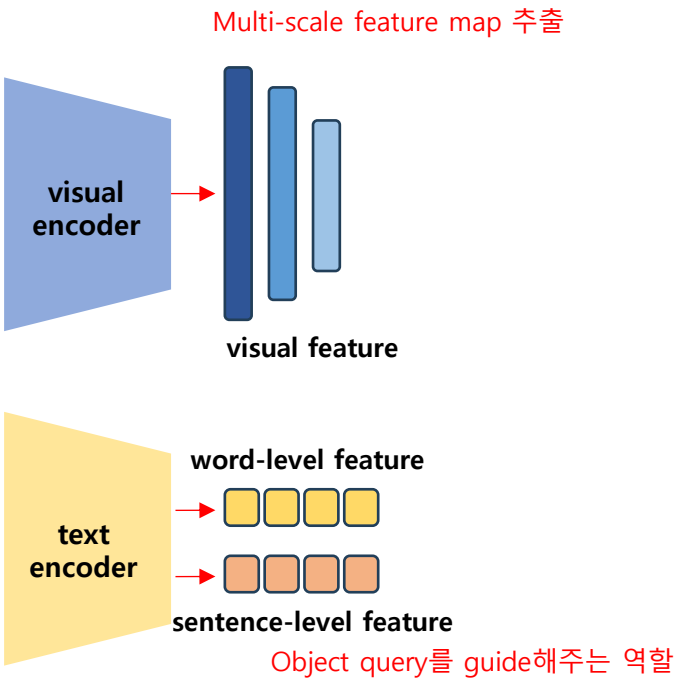
Referformer

Language as Queries for Referring Video Object Segmentation

❖ Methods

- 전체 파이프라인: **Backbone** → Transformer → CM-FPN→ Segmentation & Instance-sequence-matching

  Visual Encoder + Text Encoder로 시각/언어 특징 추출



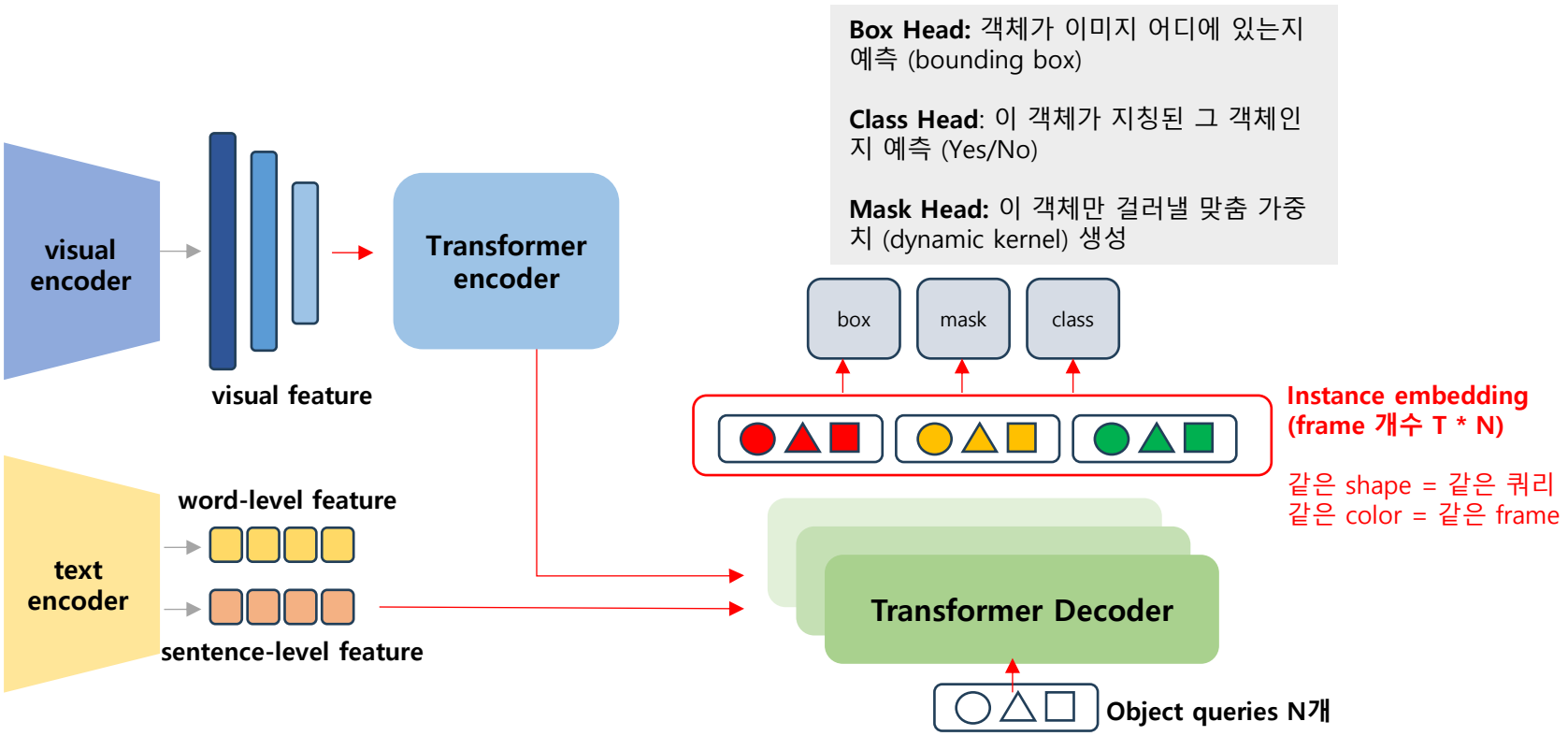
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Methods

- 전체 파이프라인: Backbone → **Transformer** → CM-FPN→ Segmentation & Instance-sequence-matching
- Conditional query를 입력으로 instance embedding 생성 → 각 Prediction head들 3가지 예측 수행
  - ➢ 각 query는 frame간 같은 weight를 공유 → 나중에 같은 객체인지 확인할 필요 X



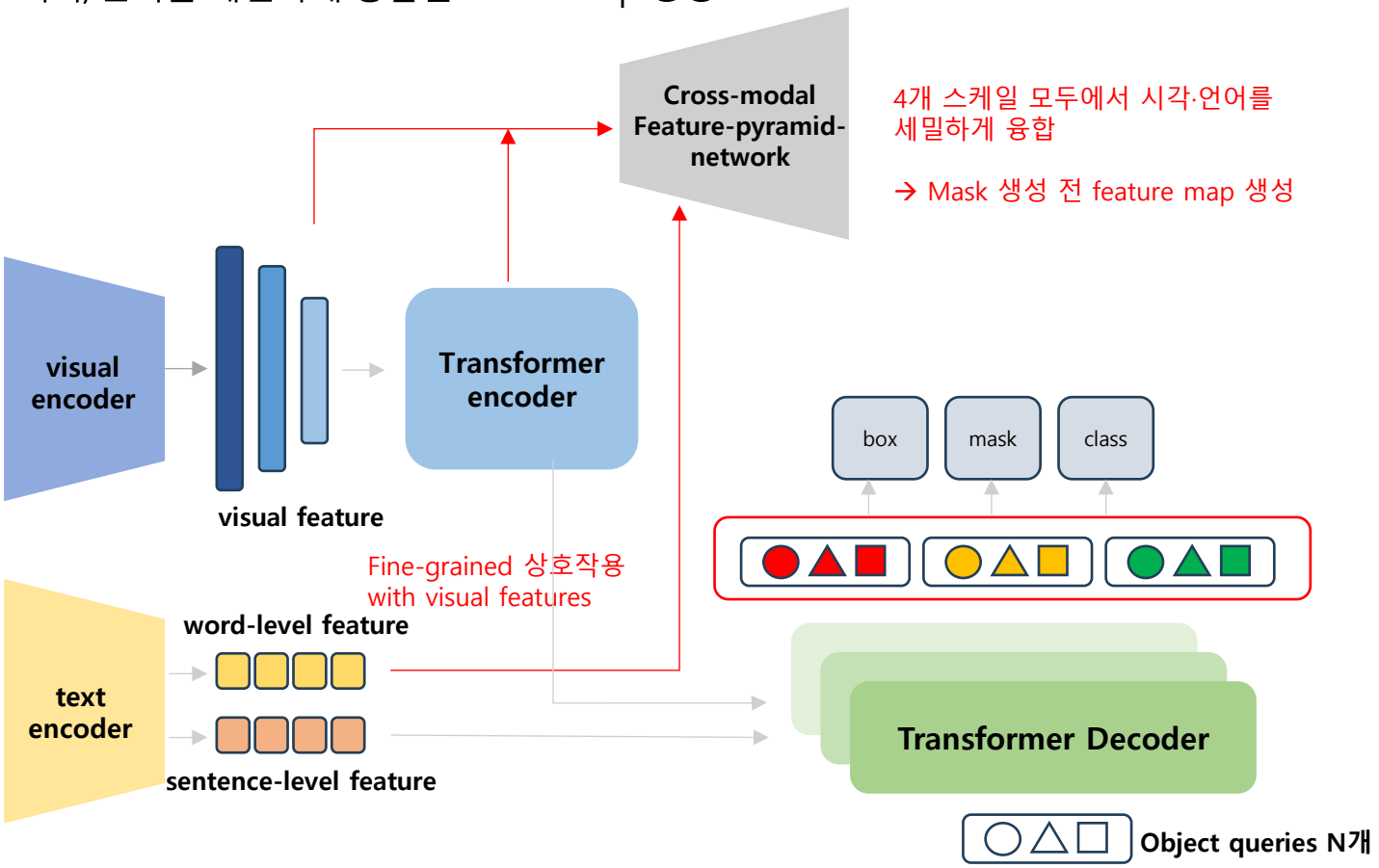
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Methods

- 전체 파이프라인: Backbone → Transformer → **CM-FPN**→ Segmentation & Instance-sequence-matching
- 시각/언어를 세밀하게 융합한 feature map 생성



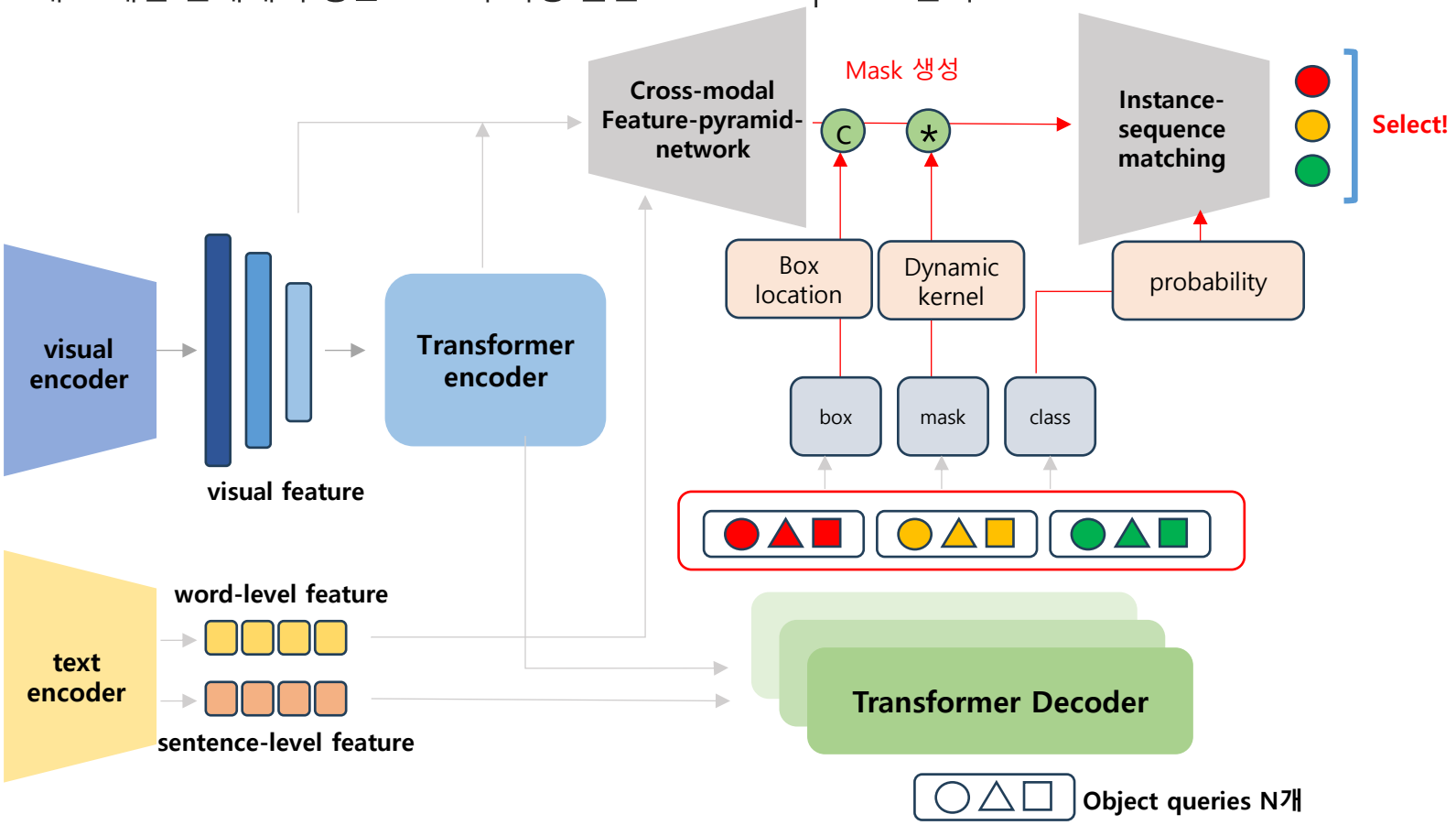
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Methods

- 전체 파이프라인: Backbone → Transformer → CM-FPN→ **Segmentation & Instance-sequence-matching**
- T개 프레임 전체에서 평균 score가 가장 높은 instance sequence 선택



고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Experiments

- Ref-Youtube-VOS, Ref-DAVIS17, A2D Sentence, JHMDB-Sentences 벤치마크에서 SOTA

Method	Backbone	Ref-Youtube-VOS			Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Spatial Visual Backbones							
CMSA [57]	ResNet-50	34.9	33.3	36.5	34.7	32.2	37.2
CMSA + RNN [57]	ResNet-50	36.4	34.8	38.1	40.2	36.9	43.5
URVOS [42]	ResNet-50	47.2	45.3	49.2	51.5	47.3	56.0
ReferFormer	ResNet-50	55.6	54.8	56.5	**58.5**	**55.8**	**61.3**
ReferFormer*	ResNet-50	**58.7**	**57.4**	**60.1**	-	-	-
PMINet [10]	ResNeSt-101	48.2	46.7	49.6	-	-	-
PMINet + CFBI [10]	ResNeSt-101	53.0	51.5	54.5	-	-	-
CITD* [23]	ResNet-101	56.4	54.8	58.1	-	-	-
ReferFormer	ResNet-101	57.3	56.1	58.4	-	-	-
ReferFormer*	ResNet-101	**59.3**	**58.1**	**60.4**	-	-	-
PMINet + CFBI [10]	Ensemble	54.2	53.0	55.5	-	-	-
CITD [23]	Ensemble	61.4	60.0	62.7	-	-	-
ReferFormer	Swin-L	62.4	60.8	64.0	**60.5**	**57.6**	**63.4**
ReferFormer*	Swin-L	**64.2**	**62.3**	**66.2**	-	-	-
Spatio-temporal Visual Backbones							
MTTR† ($\omega = 12$) [3]	Video-Swin-T	55.3	54.0	56.6	-	-	-
ReferFormer † ($\omega = 5$)		**56.0**	**54.8**	**57.3**	-	-	-
ReferFormer	Video-Swin-T	59.4	58.0	60.9	-	-	-
ReferFormer*		**62.6**	**59.9**	**63.3**	-	-	-
ReferFormer	Video-Swin-S	60.1	58.6	61.6	-	-	-
ReferFormer*		**63.3**	**61.4**	**65.2**	-	-	-
ReferFormer	Video-Swin-B	62.9	61.3	64.6	**61.1**	**58.1**	**64.1**
ReferFormer*		**64.9**	**62.8**	**67.0**	-	-	-

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu *et al.* [15]	VGG-16	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
Gavrilyuk *et al.* [13]	I3D	47.5	34.7	21.1	8.0	0.2	53.6	42.1	19.8
CMSA + CFSA [58]	ResNet-101	48.7	43.1	35.8	23.1	5.2	61.8	43.2	-
ACAN [48]	I3D	55.7	45.9	31.9	16.0	2.0	60.1	49.0	27.4
CMPC-V [28]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
ClawCraneNet [22]	ResNet-50/101	70.4	67.7	61.7	48.9	17.1	63.1	59.9	-
MTTR ($\omega = 8$) [3]	Video-Swin-T	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
MTTR ($\omega = 10$) [3]	Video-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer† ($\omega = 6$)	Video-Swin-T	76.0	72.2	65.4	49.8	17.9	72.3	64.1	48.6
ReferFormer ($\omega = 5$)	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
ReferFormer ($\omega = 5$)	Video-Swin-S	82.6	79.4	73.1	57.4	21.1	77.7	69.8	53.9
ReferFormer ($\omega = 5$)	Video-Swin-B	**83.1**	**80.4**	**74.1**	**57.9**	**21.2**	**78.6**	**70.3**	**55.0**

Table 2. Comparison with the state-of-the-art methods on A2D-Sentences. † means our model is trained from scratch.

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu *et al.* [15]	VGG-16	63.3	35.0	8.5	0.2	0.0	54.6	52.8	17.8
Gavrilyuk *et al.* [13]	I3D	69.9	46.0	17.3	1.4	0.0	54.1	54.2	23.3
CMSA + CFSA [58]	ResNet-101	76.4	62.5	38.9	9.0	0.1	62.8	58.1	-
ACAN [48]	I3D	75.6	56.4	28.7	3.4	0.0	57.6	58.4	28.9
CMPC-V [28]	I3D	81.3	65.7	37.1	7.0	0.0	61.6	61.7	34.2
ClawCraneNet [22]	ResNet-50/101	88.0	79.6	56.6	14.7	0.2	64.4	65.6	-
MTTR ($\omega = 8$) [3]	Video-Swin-T	91.0	81.5	57.0	14.4	0.1	67.4	67.9	36.6
MTTR ($\omega = 10$) [3]	Video-Swin-T	93.9	85.2	61.6	16.6	0.1	70.1	69.8	39.2
ReferFormer† ($\omega = 6$)	Video-Swin-T	93.3	84.2	61.4	16.4	0.3	70.0	69.3	39.1
ReferFormer ($\omega = 5$)	Video-Swin-T	95.8	89.3	66.8	18.9	0.2	71.9	71.0	42.2
ReferFormer ($\omega = 5$)	Video-Swin-S	95.8	90.1	68.7	20.3	0.2	72.8	71.5	42.4
ReferFormer ($\omega = 5$)	Video-Swin-B	**96.2**	**90.2**	**70.2**	**21.0**	**0.3**	**73.0**	**71.8**	**43.7**

Table 3. Comparison with the state-of-the-art methods on JHMDB-Sentences. † means our model is trained from scratch.

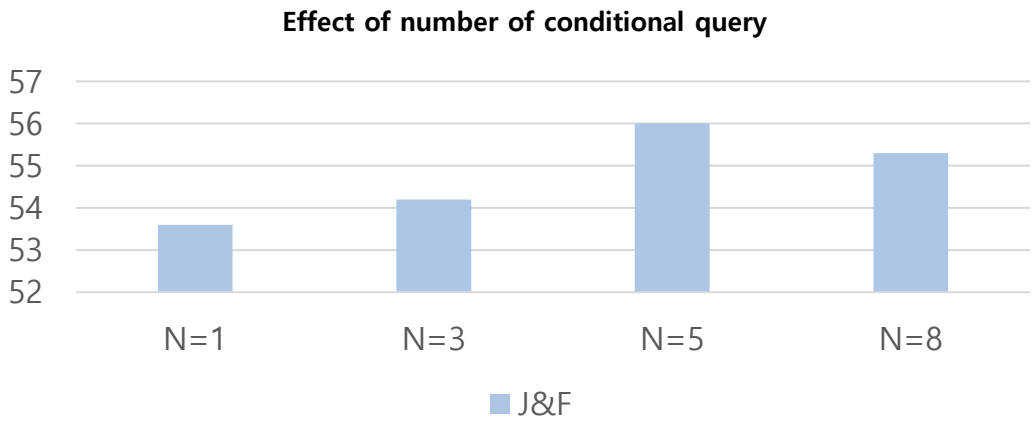
고려대학교 KOREA UNIVERSITY

Referformer

Language as Queries for Referring Video Object Segmentation

❖ Experiments

- Conditional query를 5개 정도 사용했을 때 최적
- 원래 DETR은 object query를 100개 정도 사용하는 것이 default → query 수를 획기적으로 줄임



고려대학교 KOREA UNIVERSITY

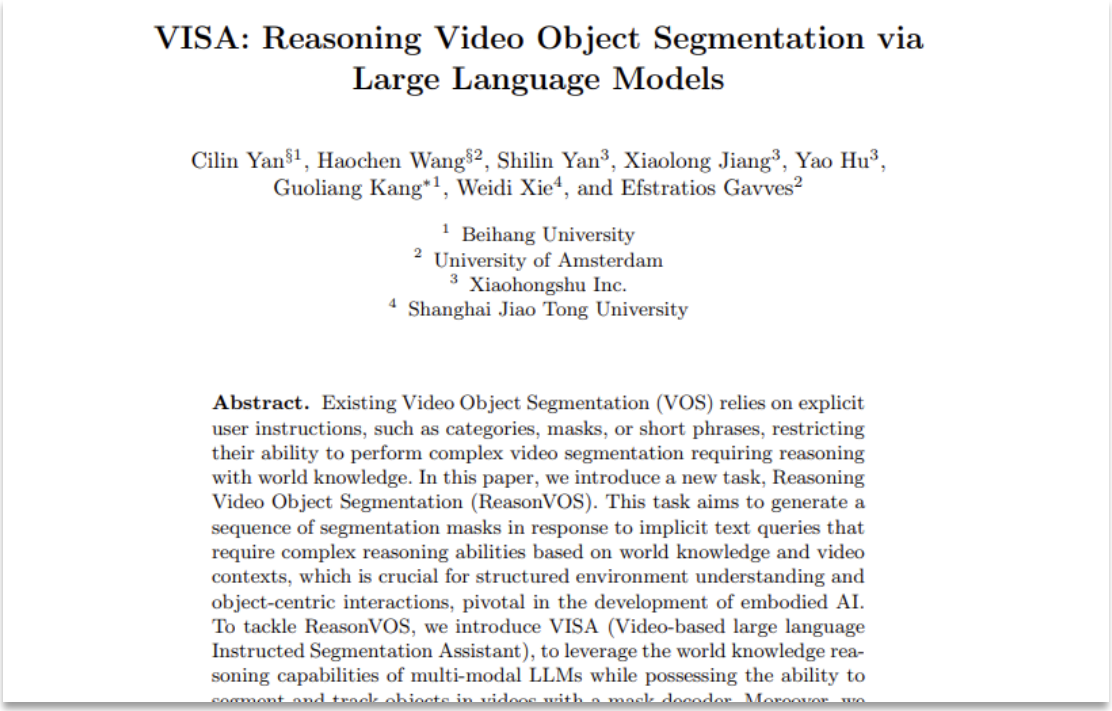
Reasoning VOS

고려대학교
KOREA UNIVERSITY

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ VISA: Reasoning Video Object Segmentation via Large Language Models

- 2024 ECCV accept poster
- 174회 인용



VISA: Reasoning Video Object Segmentation via Large Language Models

Cilin Yan[§1], Haochen Wang[§2], Shilin Yan[3], Xiaolong Jiang[3], Yao Hu[3], Guoliang Kang[*1], Weidi Xie[4], and Efstratios Gavves[2]

[1] Beihang University
[2] University of Amsterdam
[3] Xiaohongshu Inc.
[4] Shanghai Jiao Tong University

Abstract. Existing Video Object Segmentation (VOS) relies on explicit user instructions, such as categories, masks, or short phrases, restricting their ability to perform complex video segmentation requiring reasoning with world knowledge. In this paper, we introduce a new task, Reasoning Video Object Segmentation (ReasonVOS). This task aims to generate a sequence of segmentation masks in response to implicit text queries that require complex reasoning abilities based on world knowledge and video contexts, which is crucial for structured environment understanding and object-centric interactions, pivotal in the development of embodied AI. To tackle ReasonVOS, we introduce VISA (Video-based large language Instructed Segmentation Assistant), to leverage the world knowledge reasoning capabilities of multi-modal LLMs while possessing the ability to segment and track objects in videos with a mask decoder. Moreover, we

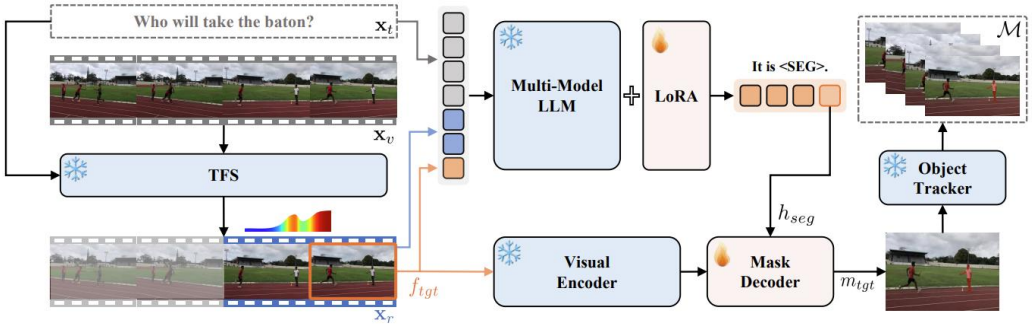
Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., & Gavves, E. (2024). VISA: Reasoning video object segmentation via large language models. *Proceedings of the European Conference on Computer Vision*.

VISA

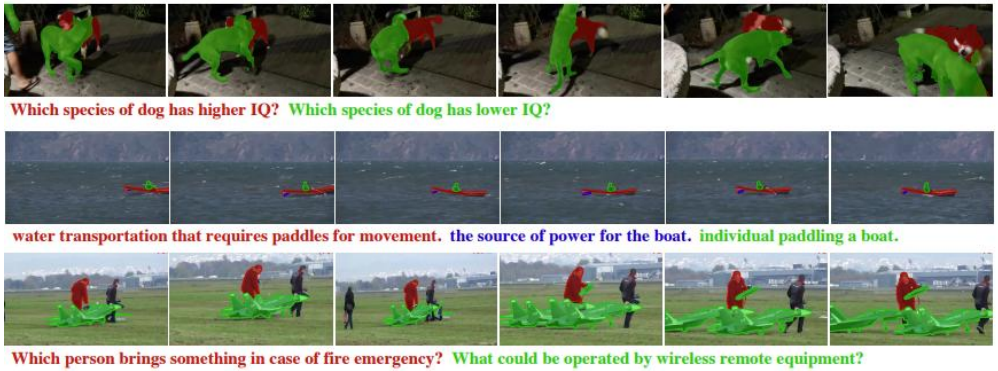
VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ VISA: Reasoning Video Object Segmentation via Large Language Models

- Implicit한 text query로 비디오 속 객체를 분할하고 추적하는 Reasoning VOS task 제안
  - ➢ Visual-language alignment만으로는 task 수행이 어려움
- 1042개의 비디오 / 35074쌍의 text query 벤치마크 ReVOS 제안



framework



ReVOS

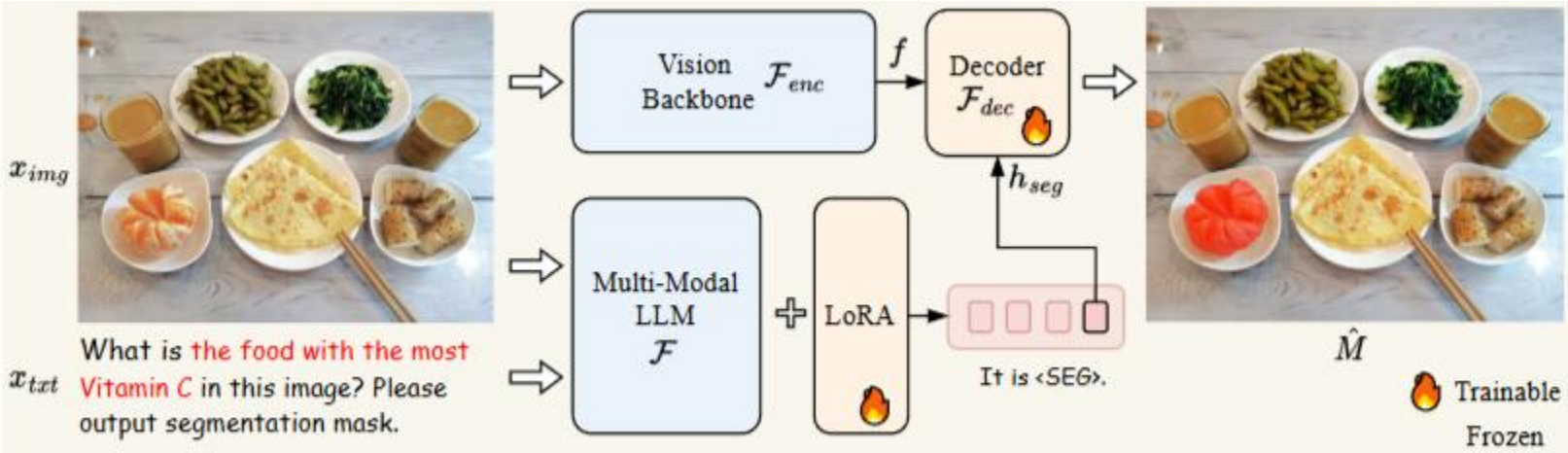
고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ VISA는 Reasoning VOS task를 어떤 방식으로 수행할까?

- Backgrounds: Image segmentation에서는 MLLM의 추론 능력을 이미 활용 (LISA)
- **Motivation: "MLLM의 추론 능력을 Video Obejct에서도 직접 활용할 수 없을까?"**



LISA(2024,CVPR)

Xia, C., Liang, J., Xiao, J., Li, Z., Loy, C. C., & Dong, C. (2024). LISA: Reasoning segmentation via large language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

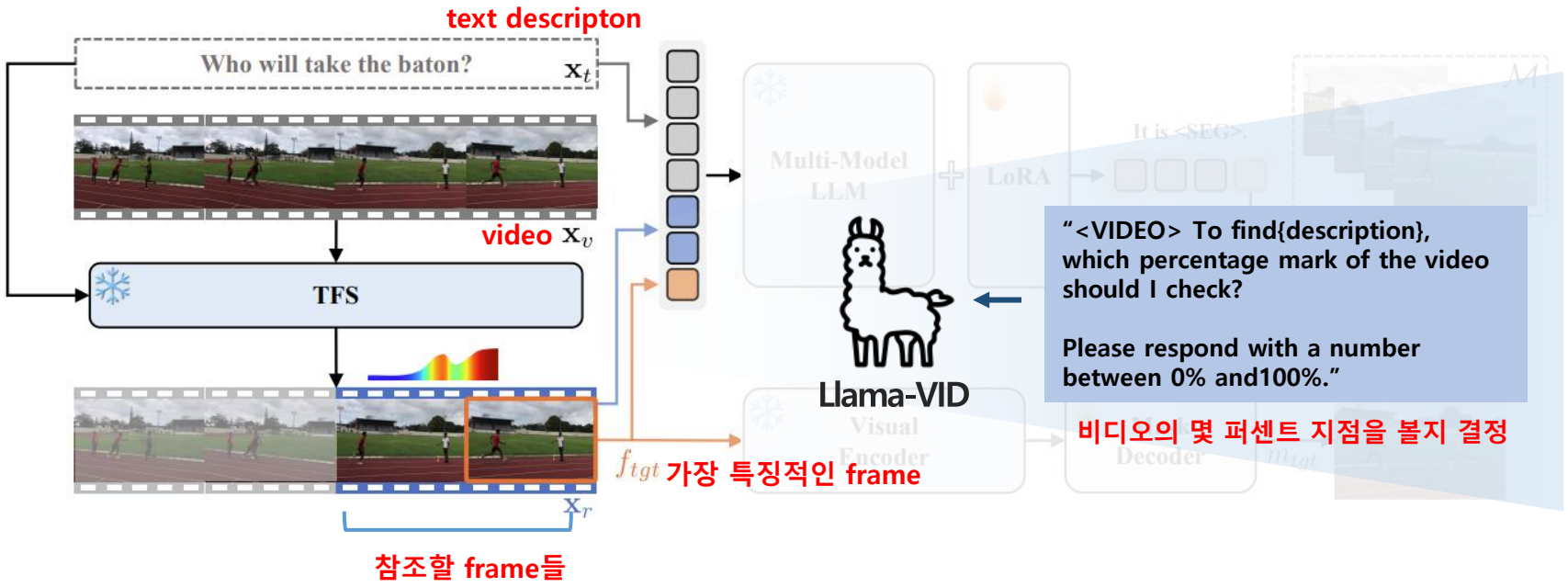
고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Methods

- 전체 파이프라인: **Text-guided Frame Sampler(TFS)** → MLLM reasoning → Mask Decoding
- VQA용 MLLM인 **LLama-VID**로 text 기반 가장 특징적인 frame f_{tgt} 과 참조할 frame들 x_r 선택
  - ➢ Llama-VID: 한 frame 당 2개의 token으로 압축해서 긴 비디오를 처리 가능, but 공간적인 정보 상실



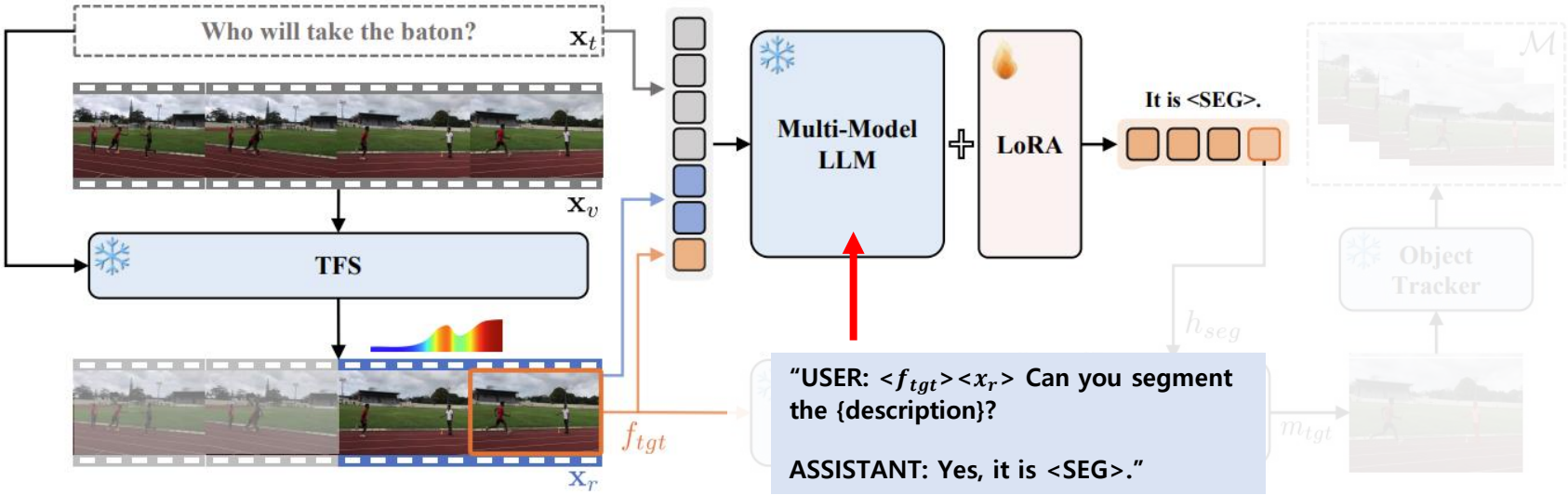
고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Methods

- 전체 파이프라인: Text-guided Frame Sampler(TFS) → **MLLM reasoning** → Mask Decoding
- 선택 frame+텍스트를 MLLM에 입력 → <SEG> token의 마지막 layer embedding 추출
- <SEG> token의 마지막 layer embedding → MLP → hidden state h_{seg} 생성



MLLM이 추론한 객체의 정체·맥락·위치 정보를 <SEG> 토큰 하나에 압축

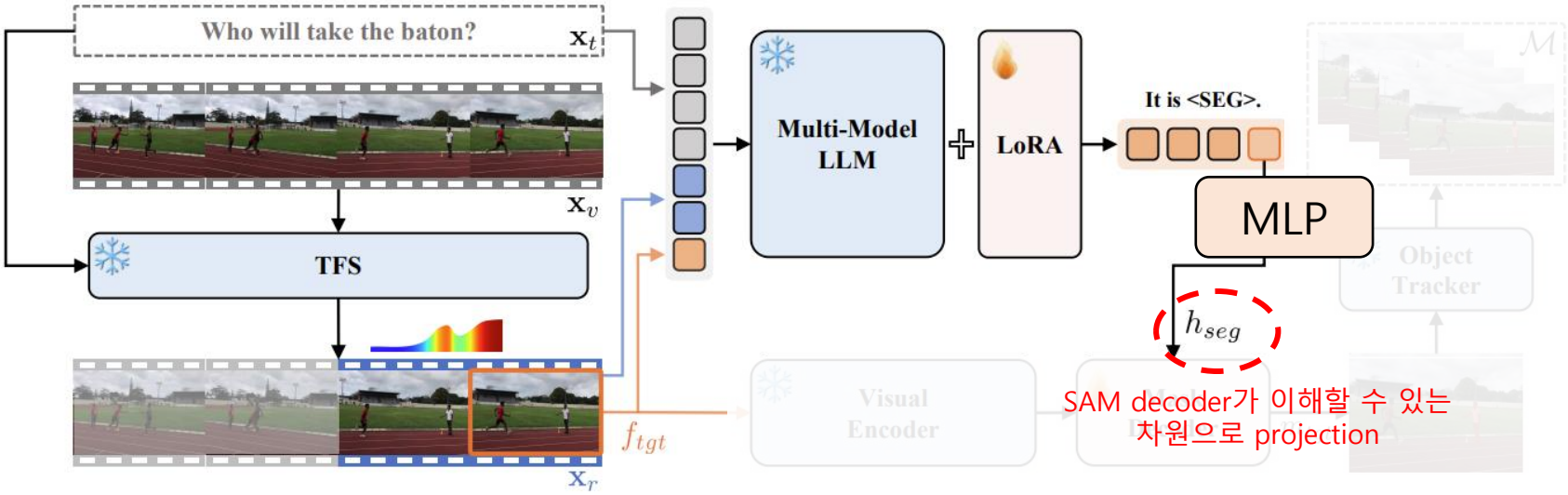
고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Methods

- 전체 파이프라인: Text-guided Frame Sampler(TFS) → **MLLM reasoning** → Mask Decoding
- 선택 frame+텍스트를 함께 입력 → <SEG> token의 마지막 layer embedding 추출
- <SEG> token의 마지막 layer embedding → MLP → hidden state h_{seg} 생성



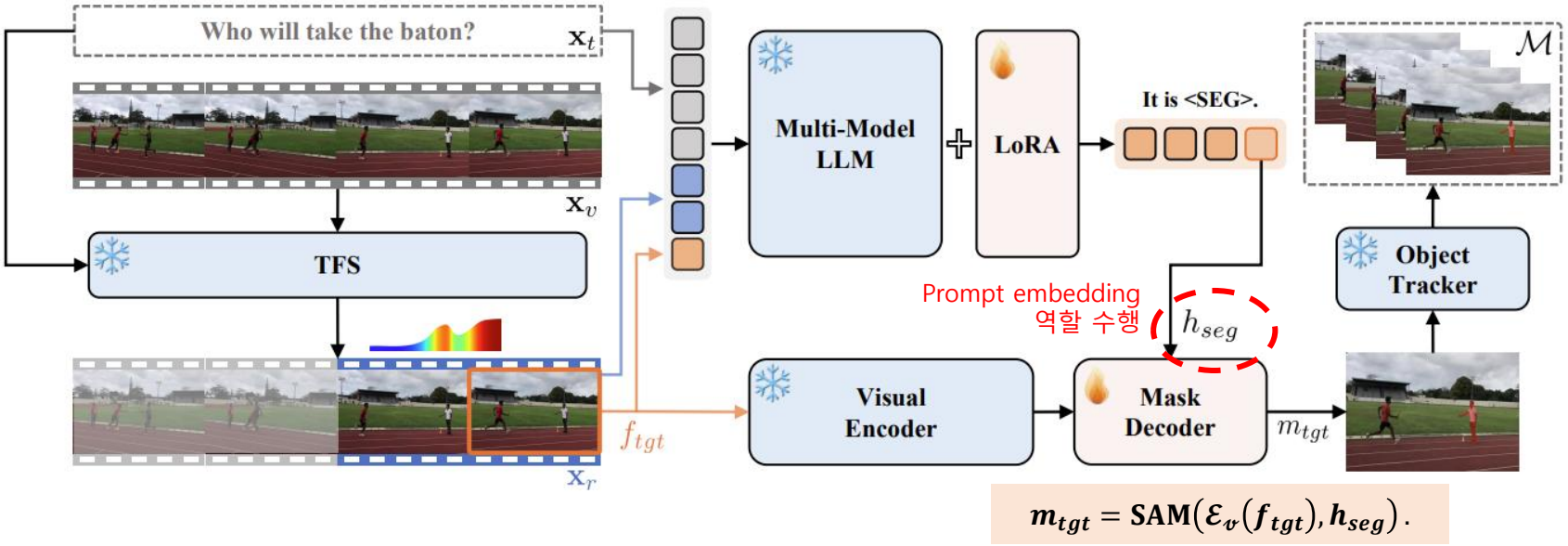
고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Methods

- 전체 파이프라인: Text-guided Frame Sampler(TFS) → MLLM reasoning → **Mask Decoding & Tracking**
- h_{seg}를 mask decoder에 넣어 keyframe의 mask 생성 → object tracker에 넣어 나머지 frame으로 전파
  - ➢ Mask Decoder: SAM(2024) / Object Tracker: XMem(2022)



고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Experiments

- 제안한 벤치마크인 ReVOS에서 SOTA 달성
- Reasoning task뿐만 아니라 referring task에서도 SOTA 달성

Table 1. Performance comparison on ReVOS dataset. $*$ means the method is reproduced in this work. (IT) means instruction tuning with the ReVOS training set. $\mathcal{R}$ is the robustness score.

Method	Backbone	referring			reasoning			overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
ReferFormer [47]	Resnet50	16.6	17.1	16.9	11.9	13.8	12.8	14.3	15.4	14.9	4.9
MTTR [2]	Video-Swin-T	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5	5.6
LMPM [8]	Swin-T	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
ReferFormer [47]	Video-Swin-B	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1	8.8
LLaMA-VID [21]+LMPM	Swin-T	29.0	39.1	34.1	12.8	23.7	18.2	20.9	31.4	26.1	3.4
LISA [17]	LLaVA-7B	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9	9.3
LISA* [17]	LLaVA-13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
TrackGPT(IT)* [38]	LLaVA-7B	46.7	49.7	48.2	36.8	41.2	39.0	41.8	45.5	43.6	11.6
TrackGPT(IT)* [38]	LLaVA-13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0	12.8
VISA	Chat-UniVi-7B	51.1	54.7	52.9	36.7	41.7	39.2	43.9	48.2	46.1	7.9
VISA	Chat-UniVi-13B	52.3	55.8	54.1	38.3	43.5	40.9	45.3	49.7	47.5	8.3
VISA(IT)	LLaVA-7B	49.4	52.6	51.0	40.5	45.8	43.2	44.9	49.2	47.1	15.3
VISA(IT)	LLaVA-13B	**55.7**	59.0	**57.4**	41.9	46.5	44.2	**48.8**	52.8	50.8	15.1
VISA(IT)	Chat-UniVi-7B	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9	**15.5**
VISA(IT)	Chat-UniVi-13B	55.6	**59.1**	**57.4**	**42.0**	**46.7**	**44.3**	**48.8**	**52.9**	**50.9**	14.5

고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Experiments

- Referring VOS 벤치마크에서도 SOTA 달성
- RVOS용 모델들보다 높은 성능 기록

Table 2: Performance comparison on Referring VOS datasets. The results on MeViS above the horizontal line are provided in LMPM [8], which are all obtained with the Swin-T backbone. The results of TrackGPT on MeVIS are generated by our reproduced model.

Methods	Backbone	MeViS			Ref-YT-VOS			Ref-DAVIS17		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS [34]	ResNet50	25.7	29.9	27.8	45.3	49.2	47.2	47.3	56.0	51.6
LBDT [11]	ResNet50	27.8	30.8	29.3	48.2	50.6	49.4	-	-	54.1
MTTR [2]	Video-Swin-T	28.8	31.2	30.0	54.0	56.6	55.3	-	-	-
ReferFormer [44]	Video-Swin-B	29.8	32.2	31.0	61.3	64.6	62.9	58.1	64.1	61.1
LMPM [8]	Swin-T	34.2	40.2	37.2	-	-	-	-	-	-
OnlineRefer [43]	Swin-L	-	-	-	**61.6**	**65.5**	**63.5**	61.6	67.7	64.8
LISA [16]	LLaVA-7B	35.1	39.4	37.2	53.4	54.3	53.9	62.2	67.3	64.8
LISA [16]	LLaVA-13B	35.8	40.0	37.9	54.0	54.8	54.4	63.2	68.8	66.0
TrackGPT [35]	LLaVA-7B	37.6	42.6	40.1	55.3	57.4	56.4	59.4	67.0	63.2
TrackGPT [35]	LLaVA-13B	39.2	43.1	41.2	58.1	60.8	59.5	62.7	70.4	66.5
VISA (Ours)	Chat-UniVi-7B	40.7	46.3	43.5	59.8	63.2	61.5	66.3	72.5	69.4
VISA (Ours)	Chat-UniVi-13B	**41.8**	**47.1**	**44.5**	61.4	64.7	63.0	**67.0**	**73.8**	**70.4**

고려대학교 KOREA UNIVERSITY

VISA

VISA: Reasoning Video Object Segmentation viaLarge Language Models

❖ Experiments

- f_{tgt} 사용했을 때 성능이 모든 세팅에서 약 2% 정도 높음 → TFS가 효과적
- Global-Local sampling을 한 reference frame을 사용했을 때 가장 성능이 높음 → 전역/지역적인 정보 잘 활용
- 참조하는 frame이 많아질 수록 성능이 향상됨

Table 5: Overall $\mathcal{J}\&\mathcal{F}$ on ReVOS with different number T_r of reference frames $\mathbf{x}_r$ and different sampling strategies.

	T_r	w/o Sample	Global	Local	Global-Local
f_0	0	42.6	-	-	-
	6	-	43.9	44.5	44.6
	12	-	44.5	44.9	45.0
f_{tgt}	0	44.3	-	-	-
	6	-	46.0	46.1	46.3
	12	-	46.7	46.3	46.9

Global: 균일하게 전체 비디오에서 sampling

Local: f_{tgt} 앞 뒤로 연속적이게 sampling

Global-Local: 절반은 global, 나머지 절반은 local sampling

f_0= 첫 frame
f_{tgt} = TFS를 통해 선택된 frame

고려대학교 KOREA UNIVERSITY

Limitations of VISA

❖ VISA의 한계는 무엇일까?

① 제한된 Temporal Context

- 단일 <SEG> 토큰 하나로 키프레임 또는 전체 비디오를 표현

  → 프레임 간 변화(inter-frame variation)와 시공간 특징을 충분히 담지 못함

② 부정확한 Keyframe Selection

- VISA는 외부 모델 LLaMA-VID로 키프레임을 선택

  → 복잡한 temporal reasoning이 필요한 비디오에서 정확한 선택이 어려움

③ 분리된 Segmentation & Propagation 모듈

- 키프레임 segmentation(SAM)과 마스크 전파(XMem)를 별개의 사전학습 모델로 처리

  → end-to-end 학습·추론 불가, 각 모듈의 오류가 독립적으로 전파

고려대학교 KOREA UNIVERSITY

❖ VISA의 한계는 무엇일까?

① 제한된 Temporal Context

➢ 단일 <SEG> 토큰 하나로 키프레임 또는 전체 비디오를 표현

① Temporal 정보를 제대로 활용하지 못하고 있다

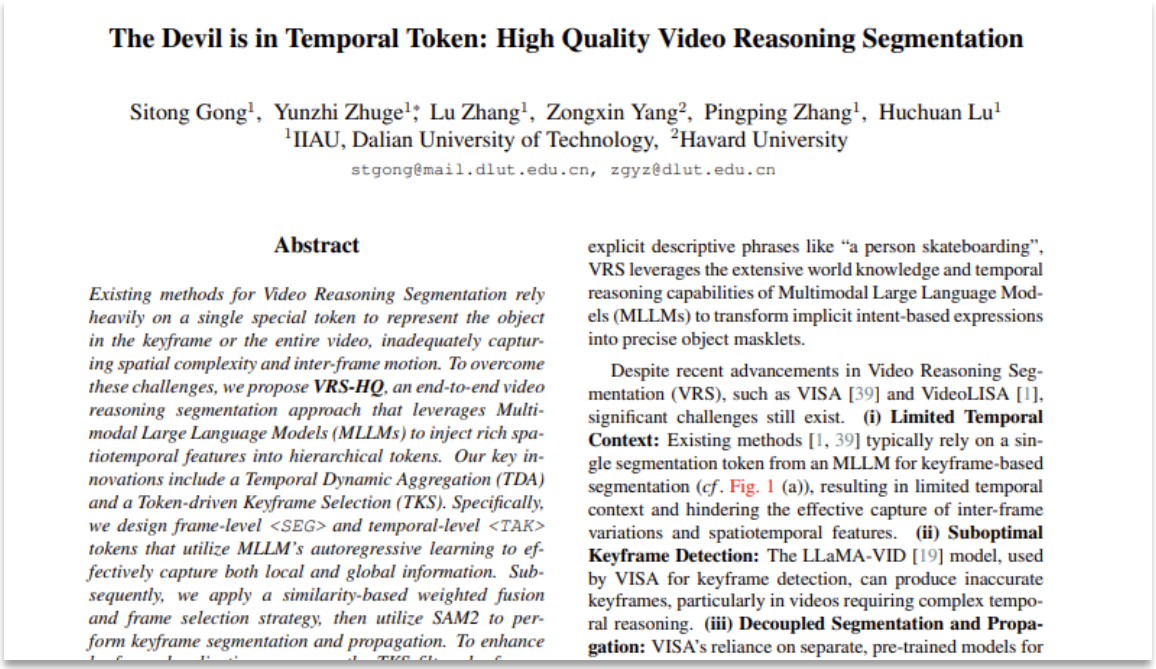
② 분리된 모듈을 이용해서 E2E 학습이 어렵다

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

- 2025 CVPR accept paper
- 52회 인용



The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

Sitong Gong[1], Yunzhi Zhuge[1]*, Lu Zhang[1], Zongxin Yang[2], Pingping Zhang[1], Huchuan Lu[1]
[1]IIAU, Dalian University of Technology, [2]Havard University
stgong@mail.dlut.edu.cn, zgyz@dlut.edu.cn

Abstract

*Existing methods for Video Reasoning Segmentation rely heavily on a single special token to represent the object in the keyframe or the entire video, inadequately capturing spatial complexity and inter-frame motion. To overcome these challenges, we propose **VRS-HQ**, an end-to-end video reasoning segmentation approach that leverages Multimodal Large Language Models (MLLMs) to inject rich spatiotemporal features into hierarchical tokens. Our key innovations include a Temporal Dynamic Aggregation (TDA) and a Token-driven Keyframe Selection (TKS). Specifically, we design frame-level <SEG> and temporal-level <TAK> tokens that utilize MLLM's autoregressive learning to effectively capture both local and global information. Subsequently, we apply a similarity-based weighted fusion and frame selection strategy, then utilize SAM2 to perform keyframe segmentation and propagation. To enhance*

explicit descriptive phrases like "a person skateboarding", VRS leverages the extensive world knowledge and temporal reasoning capabilities of Multimodal Large Language Models (MLLMs) to transform implicit intent-based expressions into precise object masklets.

Despite recent advancements in Video Reasoning Segmentation (VRS), such as VISA [39] and VideoLISA [1], significant challenges still exist. **(i) Limited Temporal Context:** Existing methods [1, 39] typically rely on a single segmentation token from an MLLM for keyframe-based segmentation (*cf*. Fig. 1 (a)), resulting in limited temporal context and hindering the effective capture of inter-frame variations and spatiotemporal features. **(ii) Suboptimal Keyframe Detection:** The LLaMA-VID [19] model, used by VISA for keyframe detection, can produce inaccurate keyframes, particularly in videos requiring complex temporal reasoning. **(iii) Decoupled Segmentation and Propagation:** VISA's reliance on separate, pre-trained models for

Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., & Lu, H. (2025). The devil is in temporal token: High quality video reasoning segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

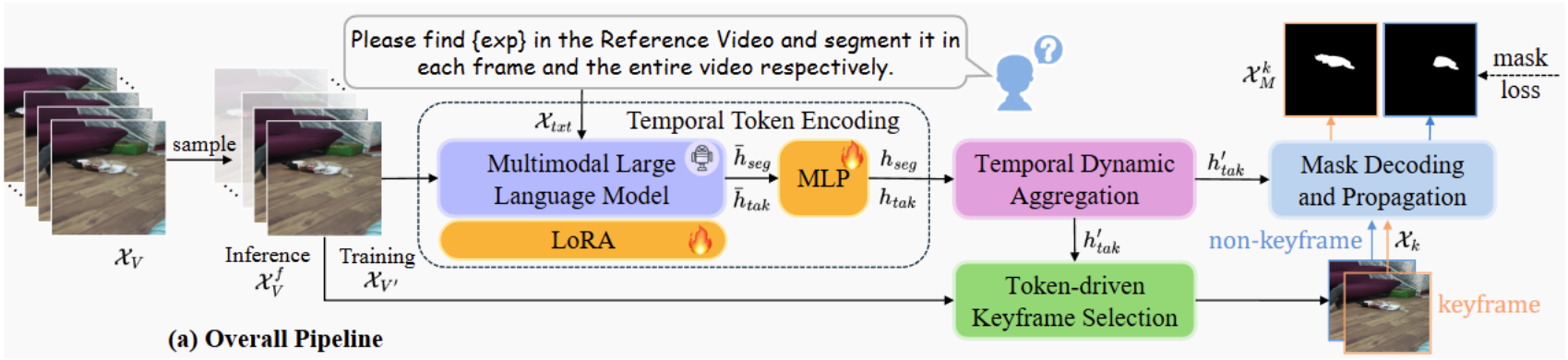
고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ VRS-HQ는 VISA의 한계를 어떻게 해결했을까?

- **Motivation 1**: 단일 토큰으로는 공간 정보와 시간 정보를 동시에 담을 수 없지 않을까?
- **Motivation 2:** SAM과 Tracker를 분리하지 말고 하나의 모델로 통합할 수 없을까?



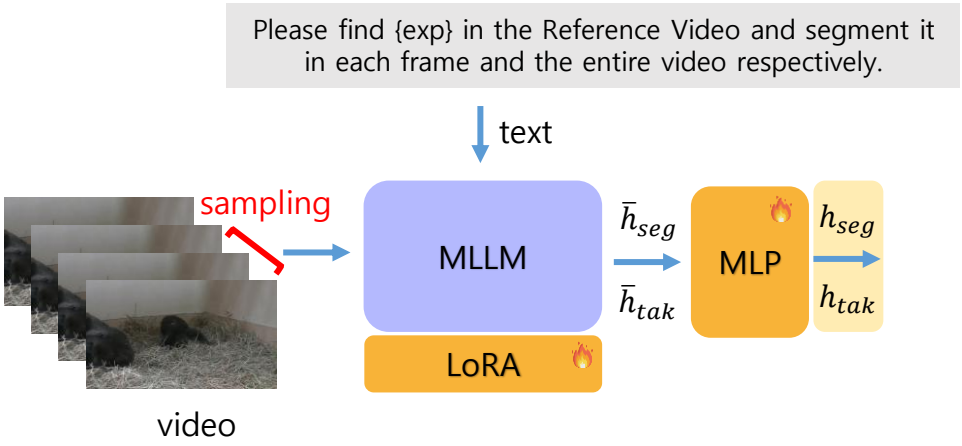
고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Methods

- 전체 파이프라인: **Temporal Token Encoding**
- CLIP 기반 Sampling video frame & text → <SEG> & <TAK> token embedding → MLP(SAM2 embedding 공간)
- <SEG>: 단일 frame의 spatial 정보 담기
- **+<TAK>:** 전체 video를 대표하는 temporal token → 프레임 간 관계 & 맥락을 통합 표현



고려대학교 KOREA UNIVERSITY

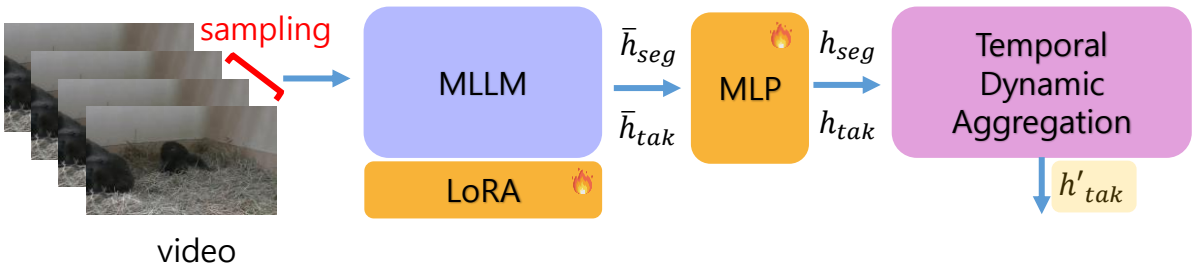
VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Methods

- 전체 파이프라인: Temporal Token Encoding → **TDA(Temporal Dynamic Aggreagation)**
- 각 frame의 <SEG>와 하나의 <TAK> cosine similarity 계산
- 정규화된 cosine similarity를 가중치로 사용해 <SEG>를 <TAK>에 합산 → 더 풍부한 시공간 표현 생성

$$h'_{tak} = h_{tak} + \alpha \sum_{i=1}^{T'} \lambda_i h_{seg}[i]$$



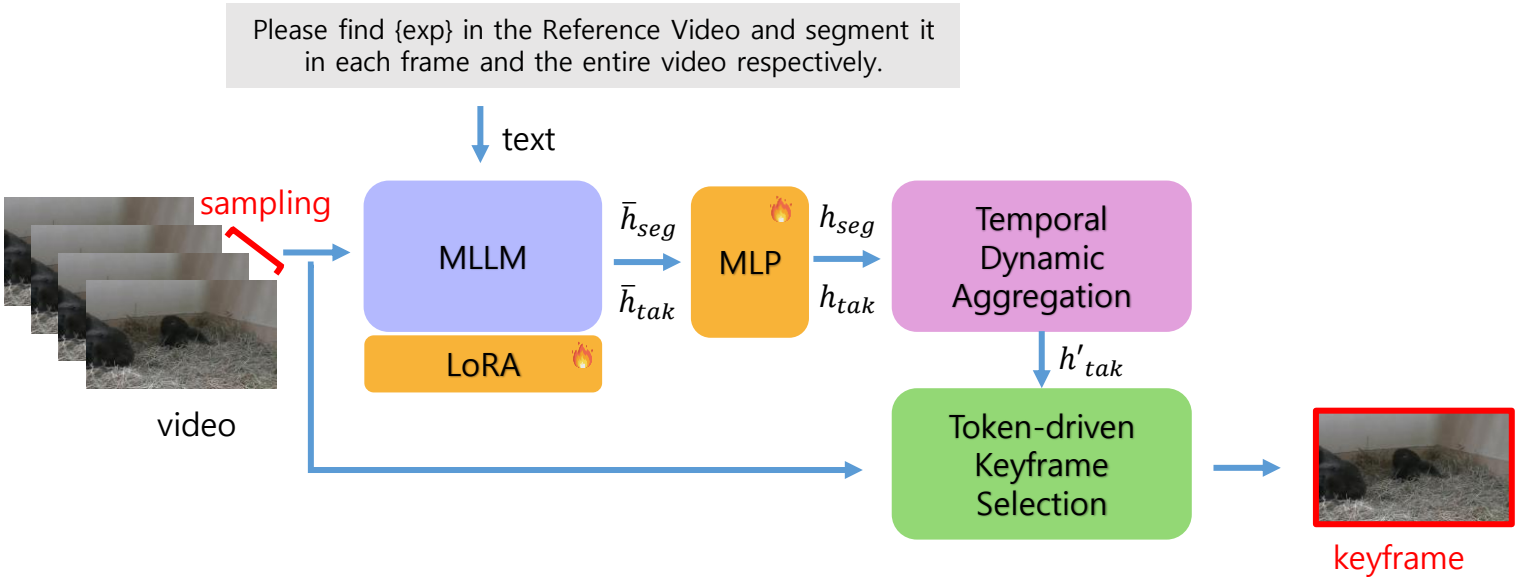
고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Methods

- 전체 파이프라인: Temporal Token Encoding → TDA → **TKS(Token-driven Keyframe Selection)**
- **VOS foundation model인 SAM2를 활용**
- Sampling video frame & h'_{tak} → SAM2에 입력 → occlusion score 산출(객체 존재 점수)
- 각 frame의 <SEG>와 <TAK>의 cosine similarity 점수랑 합산해서 keyframe 선택



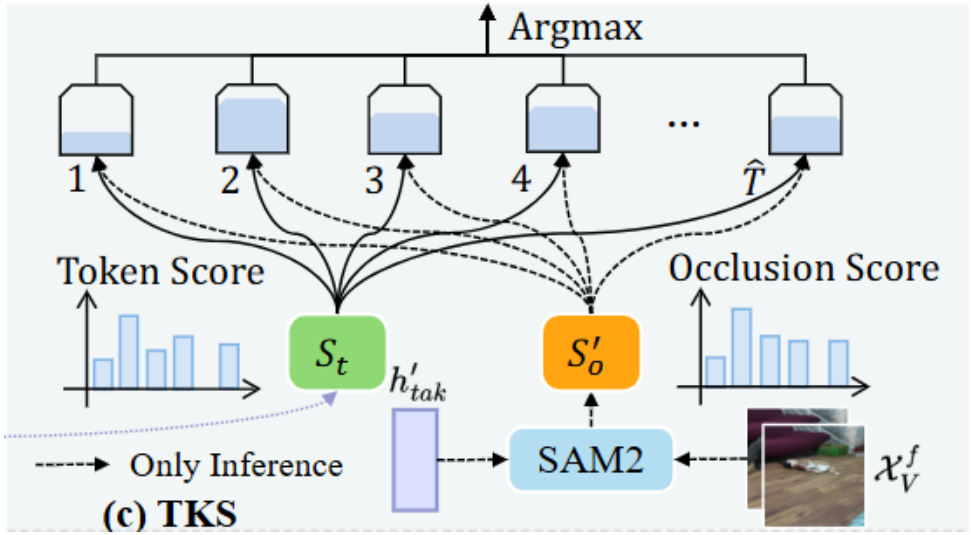
고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Methods

- 전체 파이프라인: Temporal Token Encoding → TDA → **TKS(Token-driven Keyframe Selection)**
- **VOS foundation model인 SAM2를 활용**
- Sampling video frame & h'_{tak} → SAM2에 입력 → occlusion score 산출(객체 존재 점수)
- 각 frame의 <SEG>와 <TAK>의 유사도를 계산한 token score와 합산해서 keyframe 선택



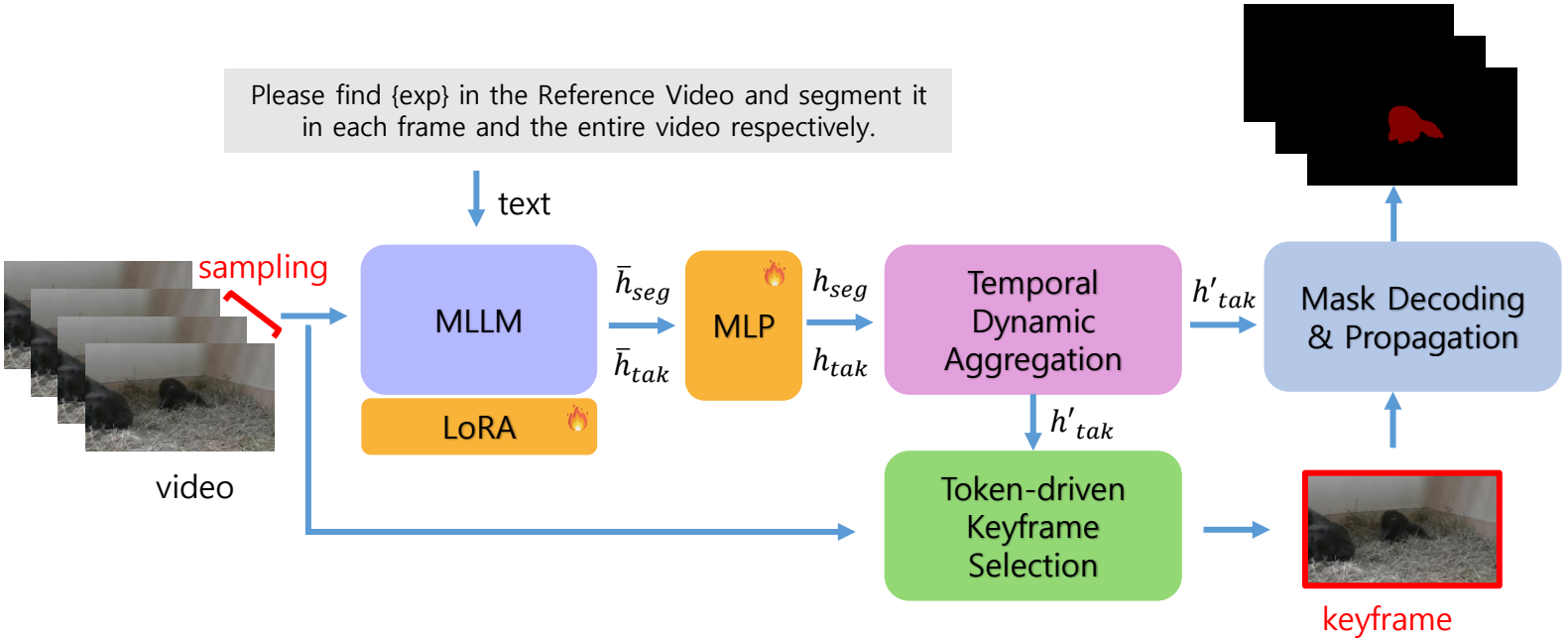
고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Methods

- 전체 파이프라인: Temporal Toekn Encoding → **TDA→ TKS → Mask Decoding & Propagation**
- 선택된 key frame mask 생성
  - ➢ 선택된 keyframe의 이미지 특징을 SAM2 image encoder로 추출 & h'_{tak} → SAM2 mask decoder
- Keyframe mask를 양방향으로 전파해서 전체 frame의 mask 생성



고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Experiments

- ReVOS와 RVOS dataset 모두에서 SOTA 기록

Table 1. Performance comparison with previous methods on ReVOS dataset.

Methods	Backbone	referring			reasoning			overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
MTTR [2] [CVPR2022]	Video-Swin-T	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5	5.6
LMPM [7] [ICCV2023]	Swin-T	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
ReferFormer [37] [CVPR2022]	Video-Swin-B	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1	8.8
LISA [17] [CVPR2024]	LLaVA-7B	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9	9.3
LISA [17] [CVPR2024]	LLaVA-13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
TrackGPT [46] [arXiv2023]	LLaVA-7B	46.7	49.7	48.2	36.8	41.2	39.0	41.8	45.5	43.6	11.6
TrackGPT [46] [arXiv2023]	LLaVA-13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0	12.8
VISA [39] [ECCV2024]	LLaVA-7B	49.4	52.6	51.0	40.5	45.8	43.2	44.9	49.2	47.1	15.3
VISA [39] [ECCV2024]	LLaVA-13B	55.7	59.0	57.4	41.9	46.5	44.2	48.8	52.8	50.8	15.1
VISA [39] [ECCV2024]	Chat-UniVi-7B	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9	15.5
VISA [39] [ECCV2024]	Chat-UniVi-13B	55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9	14.5
VRS-HQ [Ours]	Chat-UniVi-7B	59.8	64.5	62.1	53.5	58.7	56.1	56.6	61.6	59.1	**19.7**
VRS-HQ [Ours]	Chat-UniVi-13B	**61.1**	**65.5**	**63.3**	**54.1**	**59.4**	**56.8**	**57.6**	**62.5**	**60.0**	18.9

Table 2. Performance comparison with previous methods on the validation sets of RVOS datasets.

Methods	Backbone	Ref-YouTube-VOS			Ref-DAVIS17			MeViS		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
MTTR [2] [CVPR2022]	Video-Swin-T	54.0	56.6	55.3	-	-	-	28.8	31.2	30.0
LMPM [7] [ICCV2023]	Swin-T	-	-	-	-	-	-	34.2	40.2	37.2
ReferFormer [37] [CVPR2022]	Video-Swin-B	61.3	64.6	62.9	58.1	64.1	61.1	29.8	32.2	31.0
OnlineRefer [36] [CVPR2023]	Swin-L	61.6	65.5	63.5	61.6	67.7	64.8	-	-	-
DsHmp [11] [CVPR2024]	Video-Swin-B	65.0	69.1	67.1	61.7	68.1	64.9	43.0	49.8	46.4
LISA [17] [CVPR2024]	LLaVA-7B	53.4	54.3	53.9	62.2	67.3	64.8	35.1	39.4	37.2
LISA [17] [CVPR2024]	LLaVA-13B	54.0	54.8	54.4	63.2	68.8	66.0	35.8	40.0	37.9
TrackGPT [46] [arXiv2023]	LLaVA-7B	55.3	57.4	56.4	59.4	67.0	63.2	37.6	42.6	40.1
TrackGPT [46] [arXiv2023]	LLaVA-13B	58.1	60.8	59.5	62.7	70.4	66.5	39.2	43.1	41.2
VISA [39] [ECCV2024]	Chat-UniVi-7B	59.8	63.2	61.5	66.3	72.5	69.4	40.7	46.3	43.5
VISA [39] [ECCV2024]	Chat-UniVi-13B	61.4	64.7	63.0	67.0	73.8	70.4	41.8	47.1	44.5
VideoLISA [1] [NeurIPS2024]	LLaVA-Phi-3-V	61.7	65.7	63.7	64.9	72.7	68.8	41.3	47.6	44.4
VRS-HQ [Ours]	Chat-UniVi-7B	68.3	72.5	70.4	**72.6**	**79.4**	**76.0**	47.6	**53.7**	50.6
VRS-HQ [Ours]	Chat-UniVi-13B	**69.0**	**73.1**	**71.0**	71.0	77.9	74.4	**48.0**	**53.7**	**50.9**

고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Experiments

- TDA+TKS 둘 다 제거 시 referring -5.3p, reasoning -4.9p

  → 두 모듈의 시너지 효과TDA 단독 기여 > TKS 단독 기여 → 시간 정보 통합이 핵심

- Fusion coefficeint
  - α=0.1이 최적 / α=0이면 TDA가 동작 안 해 성능 급락
  - α 너무 크면 프레임 노이즈가 <TAK>에 과도하게 주입 → 오히려 성능 저하

Table 4. Ablation analysis of TDA and TKS components.

Components	referring			reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o TDA + TKS	54.5	59.1	56.8	48.6	53.9	51.2
w/o TDA	56.5	61.1	58.8	51.5	56.7	54.1
w/o TKS	57.8	62.4	60.1	50.9	56.2	53.5
VRS-HQ	**59.8**	**64.5**	**62.1**	**53.5**	**58.7**	**56.1**

Table 5. Ablation analysis of the fusion coefficient α.

α	referring			reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
0	56.5	61.1	58.8	51.5	56.7	54.1
0.1	**59.8**	**64.5**	**62.1**	**53.5**	**58.7**	**56.1**
0.25	58.9	63.8	61.3	52.3	57.8	55.0
0.5	58.2	63.0	60.6	51.4	56.6	54.0

고려대학교 KOREA UNIVERSITY

VRS-HQ

The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

❖ Experiments

- occlusion score(S3)가 가장 큰 영향력을 보임
  - ➢ referring +2.3p, reasoning +1.7p
  - ➢ 세 점수 모두 사용 시 최고 성능

- Sampling 전략 중에는 CLIP sampling이 가장 성능이 좋음

Table 6. Ablation analysis of Token-driven Keyframe Selection. S_1, S_2, and S_3 represent CLIP scores, token similarity scores, and occlusion scores, respectively.

S_1	S_2	S_3	referring			reasoning		
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
√	√	√	59.7	64.2	61.9	52.5	57.8	55.2
√	√		57.8	62.4	60.1	50.9	56.2	53.6
√		√	59.6	64.1	61.8	52.5	57.7	55.1
	√	√	**59.8**	**64.5**	**62.1**	**53.5**	**58.7**	**56.1**

Table 9. Ablation analysis of sampling strategy.

Sampling strategy	referring			reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Random Sampling	59.0	63.6	61.3	53.2	58.4	55.8
Uniform Sampling	59.3	63.9	61.6	53.3	58.4	55.8
CLIP Sampling	**59.8**	**64.5**	**62.1**	**53.5**	**58.7**	**56.1**

고려대학교 KOREA UNIVERSITY

Conclusion

고려대학교
KOREA UNIVERSITY

Conclusion

Referring to Reasoning

1. **ReferFormer (2022, CVPR)**
   - Language as Queries 패러다임 최초 제안: object query에 언어 가이드라인를 직접 주입
     → 처음부터 지칭하는 객체에 집중해 탐색
   - 탐지/분할/추적을 end-to-end 통합

2. **VISA (2024, ECCV)**
   - 단일 <SEG> token으로 MLLM 추론과 분할 연결
   - 제한된 시공간 표현력과 mask segmentation과 tracking이 분리된 구조 한계

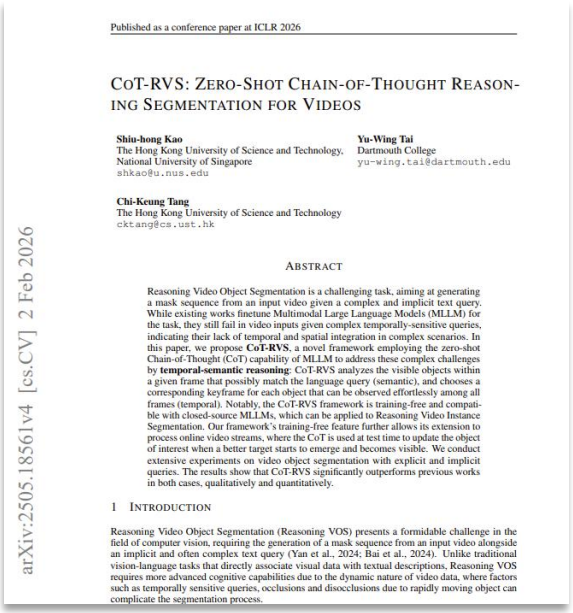
3. **VRS-HQ (2025, CVPR)**
   - <SEG>로 각 frame의 공간 정보를, <TAK>으로 전체 video의 시간 정보
     → 계층적 시공간 표현 제안
   - SAM2 도입으로 end-to-end 학습·추론 달성
   - VISA의 한계점 극복

고려대학교 KOREA UNIVERSITY

New Trend of Reasoning VOS

❖ **Training-free, Zero-shot 방향의 연구들이 활발히 진행 중**

- CoT-RVS(ICLR, 2026)
- Refer-Agent(arxiv, 2026)
- AgentRVOS(arxiv, 2026)



Published as a conference paper at ICLR 2026

CoT-RVS: Zero-Shot Chain-of-Thought Reasoning Segmentation for Videos

Shiu-hong Kao
The Hong Kong University of Science and Technology, National University of Singapore
shkao@u.nus.edu

Yu-Wing Tai
Dartmouth College
yu-wing.tai@dartmouth.edu

Chi-Keung Tang
The Hong Kong University of Science and Technology
cktang@cs.ust.hk

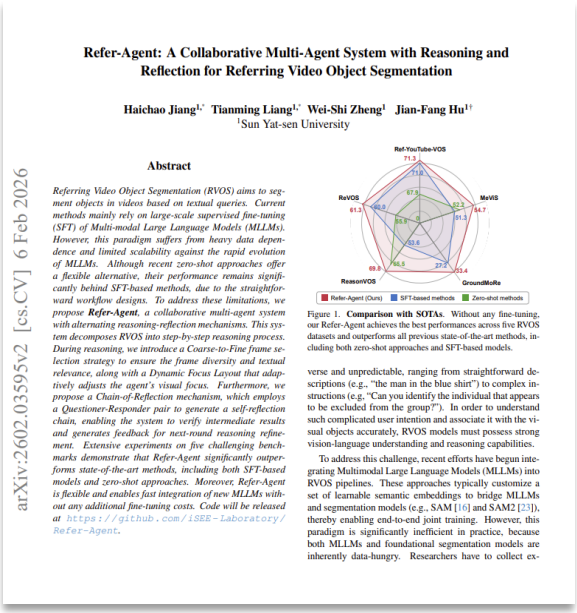
Abstract

Reasoning Video Object Segmentation is a challenging task, aiming at generating a mask sequence from an input video given a complex and implicit text query. While existing works finetune Multimodal Large Language Models (MLLM) for the task, they still fail in video inputs given complex temporally-sensitive queries, indicating their lack of temporal and spatial integration in complex scenarios. In this paper, we propose **CoT-RVS**, a novel framework employing the zero-shot Chain-of-Thought (CoT) capability of MLLM to address these complex challenges by **temporal-semantic reasoning**: CoT-RVS analyzes the visible objects within a given frame that possibly match the language query (semantic), and chooses a corresponding keyframe for each object that can be observed effortlessly among all frames (temporal). Notably, the CoT-RVS framework is training-free and compatible with closed-source MLLMs, which can be applied to Reasoning Video Instance Segmentation. Our framework's training-free feature further allows its extension to process online video streams, where the CoT is used at test time to update the object of interest when a better target starts to emerge and becomes visible. We conduct extensive experiments on video object segmentation with explicit and implicit queries. The results show that CoT-RVS significantly outperforms previous works in both cases, qualitatively and quantitatively.

1 Introduction

Reasoning Video Object Segmentation (Reasoning VOS) presents a formidable challenge in the field of computer vision, requiring the generation of a mask sequence from an input video alongside an implicit and often complex text query (Yan et al., 2024; Bai et al., 2024). Unlike traditional vision-language tasks that directly associate visual data with textual descriptions, Reasoning VOS requires more advanced cognitive capabilities due to the dynamic nature of video data, where factors such as temporally sensitive queries, occlusions and disocclusions due to rapidly moving object can complicate the segmentation process.

arXiv:2505.18561v4 [cs.CV] 2 Feb 2026



Refer-Agent: A Collaborative Multi-Agent System with Reasoning and Reflection for Referring Video Object Segmentation

Haichao Jiang[1,*] Tianming Liang[1,*] Wei-Shi Zheng[1] Jian-Fang Hu[1,†]
[1]Sun Yat-sen University

Abstract

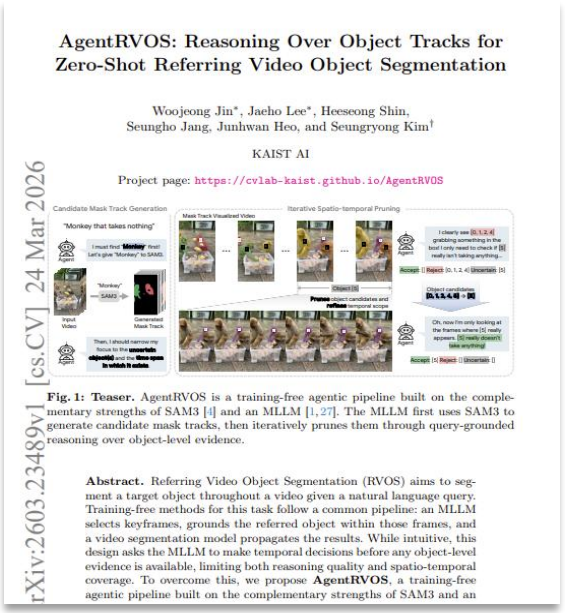
*Referring Video Object Segmentation (RVOS) aims to segment objects in videos based on textual queries. Current methods mainly rely on large-scale supervised fine-tuning (SFT) of Multi-modal Large Language Models (MLLMs). However, this paradigm suffers from heavy data dependence and limited scalability against the rapid evolution of MLLMs. Although recent zero-shot approaches offer a flexible alternative, their performance remains significantly behind SFT-based methods, due to the straightforward workflow designs. To address these limitations, we propose **Refer-Agent**, a collaborative multi-agent system with alternating reasoning-reflection mechanisms. This system decomposes RVOS into step-by-step reasoning process. During reasoning, we introduce a Coarse-to-Fine frame selection strategy to ensure the frame diversity and textual relevance, along with a Dynamic Focus Layout that adaptively adjusts the agent's visual focus. Furthermore, we propose a Chain-of-Reflection mechanism, which employs a Questioner-Responder pair to generate a self-reflection chain, enabling the system to verify intermediate results and generates feedback for next-round reasoning refinement. Extensive experiments on five challenging benchmarks demonstrate that Refer-Agent significantly outperforms state-of-the-art methods, including both SFT-based models and zero-shot approaches. Moreover, Refer-Agent is flexible and enables fast integration of new MLLMs without any additional fine-tuning costs. Code will be released at https://github.com/iSEE-Laboratory/Refer-Agent.*

Figure 1. **Comparison with SOTAs**. Without any fine-tuning, our Refer-Agent achieves the best performances across five RVOS datasets and outperforms all previous state-of-the-art methods, including both zero-shot approaches and SFT-based models.

verse and unpredictable, ranging from straightforward descriptions (e.g., "the man in the blue shirt") to complex instructions (e.g, "Can you identify the individual that appears to be excluded from the group?"). In order to understand such complicated user intention and associate it with the visual objects accurately, RVOS models must possess strong vision-language understanding and reasoning capabilities.

To address this challenge, recent efforts have begun integrating Multimodal Large Language Models (MLLMs) into RVOS pipelines. These approaches typically customize a set of learnable semantic embeddings to bridge MLLMs and segmentation models (e.g., SAM [16] and SAM2 [23]), thereby enabling end-to-end joint training. However, this paradigm is significantly inefficient in practice, because both MLLMs and foundational segmentation models are inherently data-hungry. Researchers have to collect ex-

arXiv:2602.03595v2 [cs.CV] 6 Feb 2026



AgentRVOS: Reasoning Over Object Tracks for Zero-Shot Referring Video Object Segmentation

Woojeong Jin*, Jaeho Lee*, Heeseong Shin,
Seungho Jang, Junhwan Heo, and Seungryong Kim†

KAIST AI

Project page: https://cvlab-kaist.github.io/AgentRVOS

Fig. 1: Teaser. AgentRVOS is a training-free agentic pipeline built on the complementary strengths of SAM3 [4] and an MLLM [1, 27]. The MLLM first uses SAM3 to generate candidate mask tracks, then iteratively prunes them through query-grounded reasoning over object-level evidence.

Abstract. Referring Video Object Segmentation (RVOS) aims to segment a target object throughout a video given a natural language query. Training-free methods for this task follow a common pipeline: an MLLM selects keyframes, grounds the referred object within those frames, and a video segmentation model propagates the results. While intuitive, this design asks the MLLM to make temporal decisions before any object-level evidence is available, limiting both reasoning quality and spatio-temporal coverage. To overcome this, we propose **AgentRVOS**, a training-free agentic pipeline built on the complementary strengths of SAM3 and an

rXiv:2603.23489v1 [cs.CV] 24 Mar 2026

고려대학교 KOREA UNIVERSITY

References

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Proceedings of the European Conference on Computer Vision*, 213–229.

[2] Wu, J., Jiang, Y., Sun, P., Yuan, Z., & Luo, P. (2022). Language as queries for referring video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4480–4490.

[3] Cheng, H. K., Tai, Y. W., & Tang, C. K. (2022). XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. *Proceedings of the European Conference on Computer Vision*, 640–658.

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollár, P., & Girshick, R. (2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

[5] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., & Jia, J. (2024). LISA: Reasoning segmentation via large language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

[6] Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C. Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

[7] Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., & Gavves, E. (2024). VISA: Reasoning video object segmentation via large language models. *Proceedings of the European Conference on Computer Vision*.

[8] Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., & Lu, H. (2025). The devil is in temporal token: High quality video reasoning segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[9] Kao, S., Tai, Y., & Tang, C. (2026). CoT-RVS: Zero-shot chain-of-thought reasoning segmentation for videos. *International Conference on Learning Representations*.

[10] Jin, W., Lee, J., Shin, H., Jang, S., Heo, J., & Kim, S. (2026). AgentRVOS: Reasoning over object tracks for zero-shot referring video object segmentation. *arXiv preprint arXiv:2603.23489*.

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1), 654.

고려대학교 KOREA UNIVERSITY

Thank you!

고려대학교
KOREA UNIVERSITY