CaRDiff: Video Salient Object Ranking Chain of Thought Reasoning for Saliency Prediction with Diffusion

Yunlong Tang[1,2,*], Gen Zhan[1], Li Yang[1], Yiting Liao[1], Chenliang Xu[2]

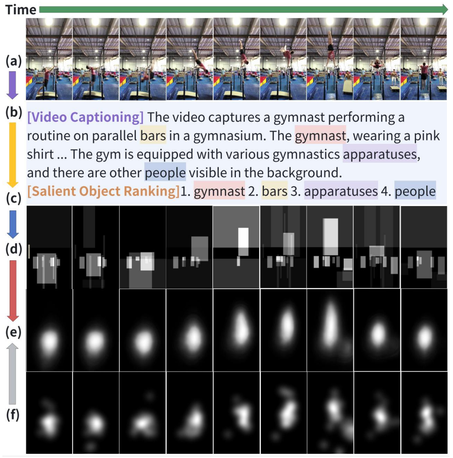
[1]ByteDance, [2]University of Rochester

*Work done during internship at ByteDance.

AAAI-25 / IAAI-25 / EAAI-25

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Contribution

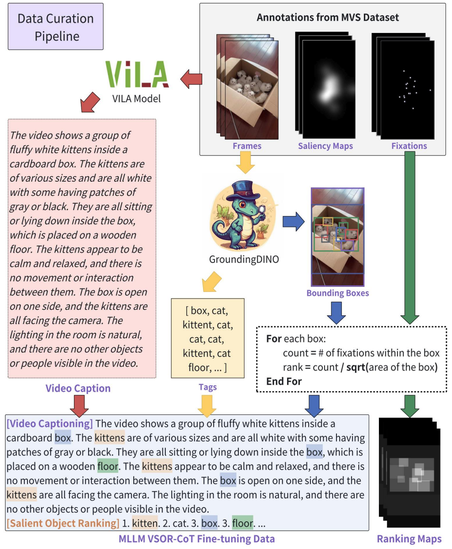


- Propose CaRDiff, an innovative video saliency prediction framework that leverages MLLM's reasoning capabilities through VSOR-CoT to analyze and rank salient objects.
- Introduce ranking maps that preserve position and ranking information of salient objects, seamlessly guiding the diffusion process to enhance saliency prediction.
- Achieve state-of-the-art performance on MVS dataset and demonstrate strong zero-shot cross-dataset capability on DHF1K benchmark.

Motivation

Traditional video saliency prediction methods ignore the role of language in visual attention guidance. We explore how language-based reasoning can enhance salient object modeling in videos.

Fine-tuning Data Curation



Dataset: Based on the MVS dataset with 1007 annotated video clips (fixation & saliency maps).
Object Detection: Using Recognize Anything and GroundingDINO to obtain bounding boxes and tags.
Saliency Ranking: Calculated using fixation maps:

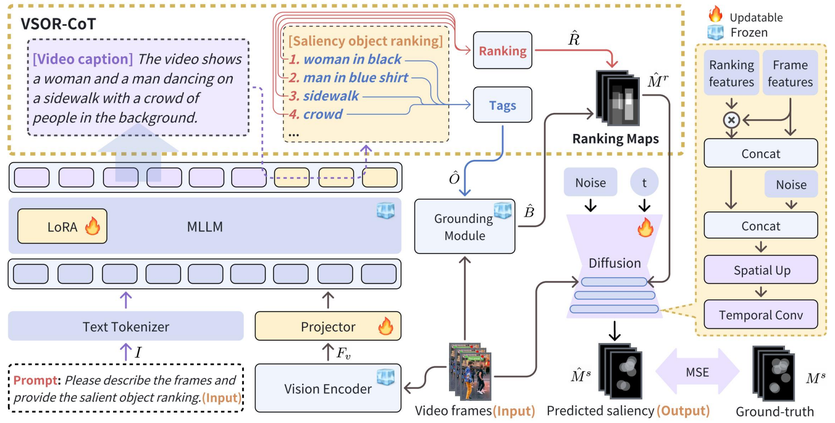
$$r_i = \frac{1}{\sqrt{|b_i|}} \sum_{(u,v)\in b_i} I[M_f(u,v) > 0]$$

where r_i reflects fixation density within the bounding box b_i.
Ranking Map Generation: Combines object positions and rankings:

$$M_r(u,v) = \sum_i r_i \cdot I[(u,v) \in b_i]$$

GT Construction: Includes video captions (via VILA-1.5) and salient object rankings for fine-tuning the MLLM with VSOR-CoT.

CaRDiff Framework

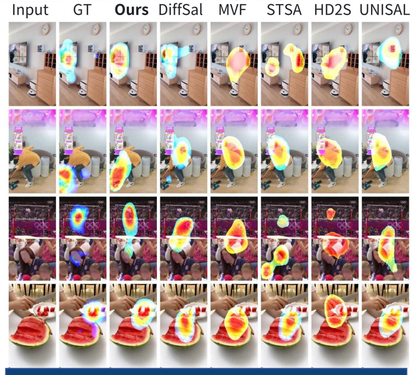


1. **VSOR-CoT**: Enhances reasoning by generating captions and salient object rankings.
2. **Grounding**: Localizes bounding boxes of object tags predicted by MLLM using GroundingDINO.
3. **Ranking Map Generation**: Converts rankings into grayscale maps. The number of object tags are determined by MLLM.
4. **Diffusion Saliency Prediction**: Predicts saliency maps via denoising diffusion.
5. **Training Pipeline**:
(1) Modality Alignment: Aligns visual features with the input space of MLLM.
(2) CoT Tuning: Fine-tunes MLLM for coherent object ranking.
(3) Diffusion Training: Optimizes saliency prediction using ranking maps.

Experiments - Main Results

Methods	Attributes				Performance			
	Video-based	Re-trained	Modality	Loss function	AUC-J	CC	Sim	NSS
ITTI (Itti, Koch, and Niebur 1998)	✗	✗	V	Non-DL	0.783	0.435	0.464	0.978
GBVS (Schölkopf, Platt, and Hofmann 2007)	✗	✗	V	Non-DL	0.808	0.492	0.491	1.097
SALICON (Huang et al. 2015)	✗	✓	V	KLD, NSS, Sim	0.814	0.523	0.512	1.261
AWS-D (Leborán et al. 2017)	✓	✗	V	Non-DL	0.675	0.240	0.384	0.560
SalGAN (Pan et al. 2017)	✗	✓	V	MSE, BCE	0.812	0.511	0.503	1.269
SAM (Cornia et al. 2018)	✗	✓	V	CC, NSS, KLD	0.818	0.531	0.522	1.274
DeepVS (Jiang et al. 2020)	✓	✗	V	KLD	0.811	0.475	0.496	1.160
ACLNet (Wang et al. 2021a)	✓	✓	V	CC, NSS, KLD	0.821	0.542	0.524	1.251
STRA-Net (Lai et al. 2020)	✓	✓	V	KLD, NSS, Sim, CC	0.826	0.563	0.531	1.289
SalEMA (Linardos et al. 2019)	✓	✓	V	BCE	0.835	0.591	0.544	1.326
TASED (Min and Corso 2019)	✓	✓	V	KLD	0.850	0.638	0.576	1.486
ESAN (Chen et al. 2021)	✓	✓	V	KLD, NSS, Sim, CC	0.853	0.645	0.590	1.517
UNISAL (Droste, Jiao, and Noble 2020)	✓	✓	V	CC, NSS, KLD	0.855	0.654	0.586	1.524
HD2S (Bellitto et al. 2021)	✓	✓	V	KLD	0.858	0.662	0.603	1.550
STSANet (Wang et al. 2021b)	✓	✓	V	KLD, CC	0.856	0.657	0.594	1.555
ViNet (Jain et al. 2021)	✓	✓	V	KLD	0.857	0.664	0.595	1.561
VSFT (Ma et al. 2022)	✓	✓	V	KLD, NSS, Sim, CC	0.857	0.666	0.597	1.572
Diff-Sal (Xiong et al. 2024)	✓	✓	V, A	MSE	0.852	0.626	0.577	1.591
MVFormer (Wen et al. 2024)	✓	✓	V	KLD, NSS, Sim	0.864	0.687	0.614	1.646
CaRDiff (ours)	✓	✓	V, L	CE, MSE	**0.870**	**0.714**	**0.630**	**1.685**

Visualization Results



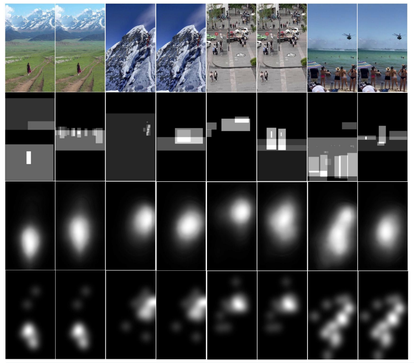
Cross-dataset Evaluation

Model	AUC-J	CC	Sim	NSS
Diff-Sal	0.802	0.218	0.192	1.069
MVFormer	0.844	0.299	0.198	1.501
CaRDiff (ours)	**0.845**	**0.312**	**0.235**	**1.584**

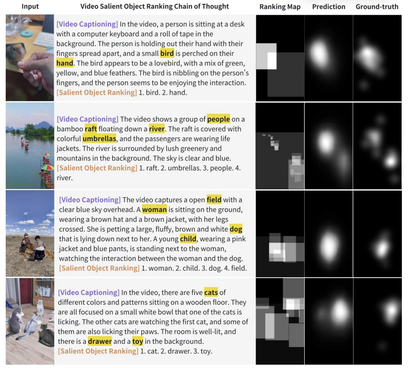
Ablation Study

Setting	AUC-J	CC	NSS	Sim
FT w/ VSOR-CoT	**0.870**	**0.714**	**1.685**	**0.630**
FT w/o VSOR-CoT	0.864	0.700	1.614	0.624
ZS w/ VSOR-CoT	0.855	0.659	1.515	0.590
ZS w/o VSOR-CoT	0.846	0.626	1.459	0.577

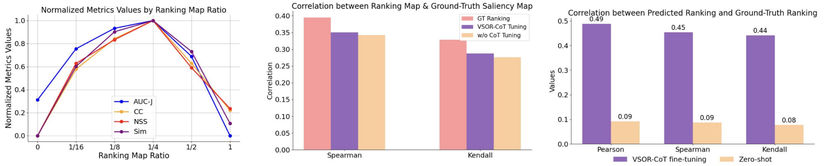
Ranking Map Replacement Experiments



More Visualization Results



Ranking Map Ratio & Correlation Experiments



Supplementary Material: https://arxiv.org/pdf/2408.12009