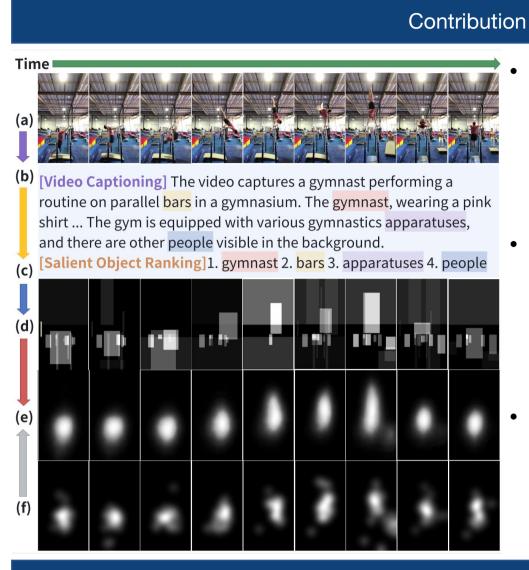


CaRDiff: Video Salient Object Ranking Chain of Thought Reasoning for Saliency Prediction with Diffusion

Yunlong $Tang^{1,2,*}$, $Gen\ Zhan^1$, $Li\ Yang^1$, $Yiting\ Liao^1$, $Chenliang\ Xu^2$ 1ByteDance , $^2University\ of\ Rochester$

*Work done during internship at ByteDance.





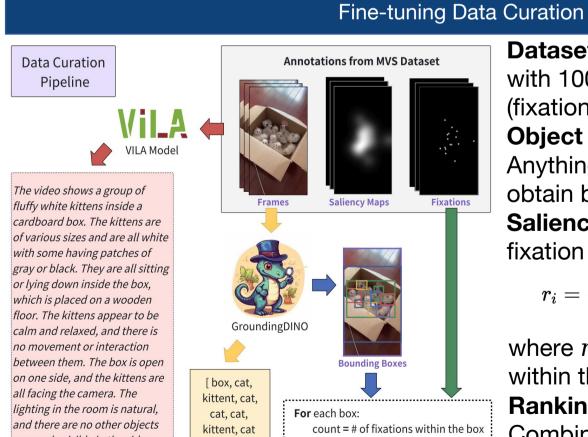
- Propose CaRDiff, an innovative video saliency prediction framework that leverages MLLM's reasoning capabilities through VSOR-CoT to analyze and rank salient objects.
- Introduce ranking maps that preserve position and ranking information of salient objects, seamlessly guiding the diffusion process 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.

rank = count / sqrt(area of the box)

End For



floor, ...]

[Video Captioning] The video shows a group of fluffy white kittens inside a cardboard box. The kittens are of various sizes and are all white with some having

patches of gray or black. They are all sitting or lying down inside the box, which is

placed on a wooden floor. The kittens appear to be calm and relaxed, and there is

no movement or interaction between them. The box is open on one side, and the kittens are all facing the camera. The lighting in the room is natural, and there are

MLLM VSOR-CoT Fine-tuning Data

or people visible in the video.

no other objects or people visible in the video.

[Salient Object Ranking] 1. kitten. 2. cat. 3. box. 3. floor

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:

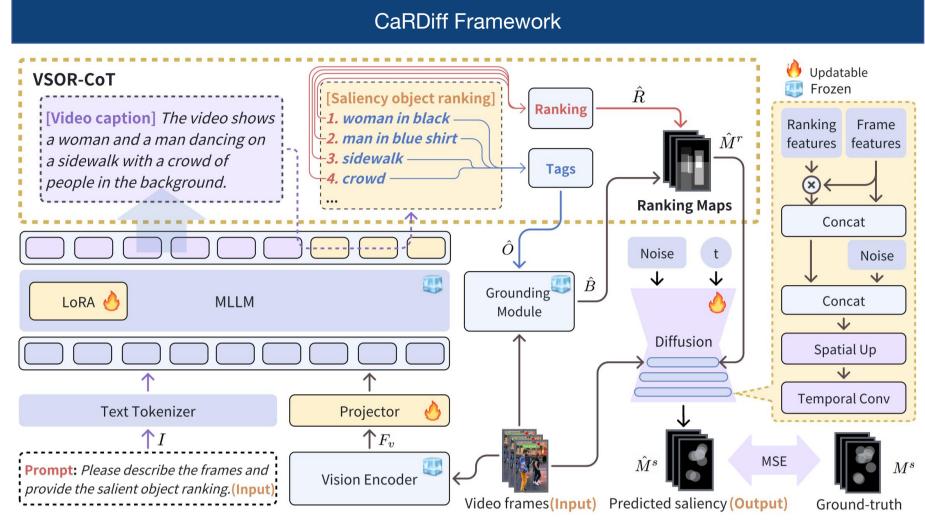
$$r_i = rac{1}{\sqrt{|b_i|}} \sum_{(u,v) \in b_i} I[M_f(u,v) > 0]$$

where r_i reflects fixation density within the bounding box b_i .

Ranking Map Generation:
Combines object positions and

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

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



- 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.

Experiments - Main Results

- 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.

	Ехро	IIIICIIIC	iviaiii	ricourto				
	Attributes				Performance			
Methods	Video- based	Re- trained	Modality	Loss function	AUC-J	CC	Sim	NSS
ITTI (Itti, Koch, and Niebur 1998)	X	X	V	Non-DL	0.783	0.435	0.464	0.978
GBVS (Schölkopf, Platt, and Hofmann 2007)	X	X	\mathbf{V}	Non-DL	0.808	0.492	0.491	1.097
SALICON (Huang et al. 2015)	X	1	\mathbf{V}	KLD, NSS, Sim	0.814	0.523	0.512	1.261
AWS-D (Leborán et al. 2017)	1	X	\mathbf{V}	Non-DL	0.675	0.240	0.384	0.560
SalGAN (Pan et al. 2017)	X	1	\mathbf{V}	MSE, BCE	0.812	0.511	0.503	1.269
SAM (Cornia et al. 2018)	X	1	\mathbf{V}	CC, NSS, KLD	0.818	0.531	0.522	1.274
DeepVS (Jiang et al. 2020)	1	X	\mathbf{V}	KLD	0.811	0.475	0.496	1.160
ACLNet (Wang et al. 2021a)	1	/	\mathbf{V}	CC, NSS, KLD	0.821	0.542	0.524	1.251
STRA-Net (Lai et al. 2020)	1	✓	\mathbf{V}	KLD, NSS, Sim, CC	0.826	0.563	0.531	1.289
SalEMA (Linardos et al. 2019)	1	1	\mathbf{V}	BCE	0.835	0.591	0.544	1.326
TASED (Min and Corso 2019)	1	1	\mathbf{V}	KLD	0.850	0.638	0.576	1.486
ESAN (Chen et al. 2021)	1	1	\mathbf{V}	KLD, NSS, Sim, CC	0.853	0.645	0.590	1.517
UNISAL (Droste, Jiao, and Noble 2020)	1	1	\mathbf{V}	CC, NSS, KLD	0.855	0.654	0.586	1.524
HD2S (Bellitto et al. 2021)	1	1	V	KLD	0.858	0.662	0.603	1.550
STSANet (Wang et al. 2021b)	1	1	\mathbf{V}	KLD, CC	0.856	0.657	0.594	1.555
ViNet (Jain et al. 2021)	1	/	\mathbf{V}	KLD	0.857	0.664	0.595	1.561
VSFT (Ma et al. 2022)	1	1	\mathbf{V}	KLD, NSS, Sim, CC	0.857	0.666	0.597	1.572
Diff-Sal (Xiong et al. 2024)	1	1	V, A	MSE	0.852	0.626	0.577	1.591
MVFormer (Wen et al. 2024)	1	1	\mathbf{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 Input GT Ours DiffSal MVF STSA HD2S UNISAL IN

FT w/ VSOR-CoT

FT w/o VSOR-CoT

ZS w/o VSOR-CoT

ZS w/ VSOR-CoT

Ranking Map Replacement Experiments

0.844

0.845

0.218

0.299

0.192

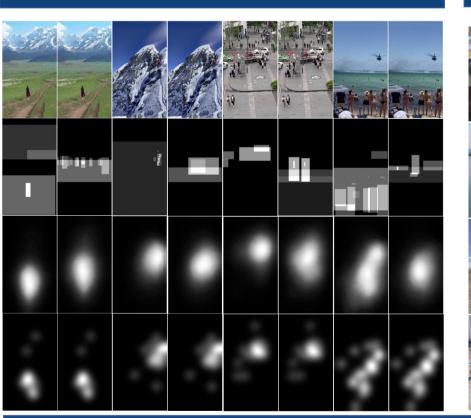
0.198

0.312 0.235 1.584

1.501

Diff-Sal

MVFormer



Input Video Salient Object Ranking Chain of Thought [Video Captioning] In the video, a person is sitting at a desk with a computer keyboard and a roll of tape in the background. The person is holding out their hand with their fingers spread apart, and a small bird is perched on their hand. The bird appears to be a lovebird, with a mix of green, yellow, and blue feathers. The bird is nibbling on the person's fingers, and the person seems to be enjoying the interaction. [Salient Object Ranking] 1. bird. 2. hand. [Video Captioning] The video shows a group of people on a bamboo raft floating down a river. The raft is covered with colorful imbrellas, and the passengers are wearing life jackets. The river is surrounded by lush greenery and mountains in the background. The sky is clear and blue. [Salient Object Ranking] 1. raft. 2. umbrellas. 3. people. 4. river. [Video Captioning] The video captures a open field with a clear blue sky overhead. A woman is sitting on the ground, wearing a brown hat and a brown jacket, with her legs crossed. She is petting a large, fluffy, brown and white dog that is lying down next to her. A young child, wearing a pink jacket and blue pants, is standing next to the woman, watching the interaction between the woman and the dog. [Salient Object Ranking] 1. woman. 2. child. 3. dog. 4. field. [Video Captioning] In the video, there are five cats of different colors and patterns sitting on a wooden floor. They are all focused on a small white bowl that one of the cats is licking. The other cats are watching the first cat, and some of them are also licking their paws. The room is well-lit, and there is a drawer and a toy in the background. [Salient Object Ranking] 1. cat. 2. drawer. 3. toy.

0.846

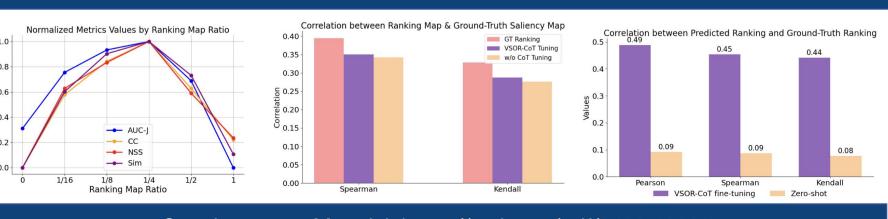
More Visualization Results

0.659

0.626

1.459

Ranking Map Ratio & Correlation Experiments



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