TF-SSD: A Strong Pipeline via Synergic Mask Filter for Training-free Co-salient Object Detection

Zhijin He[1*] **Shuo Jin**[1,2*] **Siyue Yu**[1†] **Shuwei Wu**[1] **Bingfeng Zhang**[3] **Li Yu**[4] **Jimin Xiao**[1]

[1]XJTLU [2]University of Liverpool [3]China University of Petroleum (East China) [4]Nanjing University of Information Science and Technology

CVPR JUNE 3-7, 2026 DENVER COLORADO

1. Introduction

- **Co-salient Object Detection (CoSOD)** finds common salient objects across a group of related images.
- **SAM** generates rich, accurate masks but lacks semantic and saliency understanding.
- **DINO** provides powerful semantic features and attention-based saliency cues.

We propose **TF-SSD**, a **training-free** framework that synergizes **SAM** (segmentation) and **DINO** (semantics).

SAM Oracle Masks (Upper Bound):

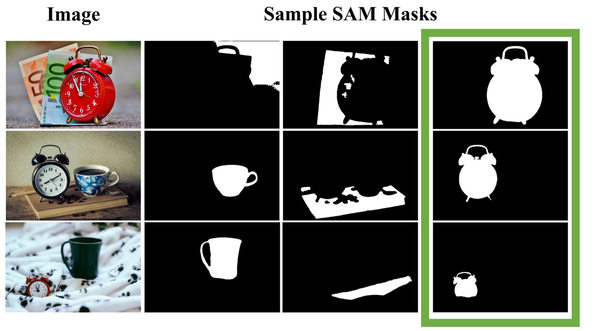

Fig. 1. SAM oracle mask examples and their upper-bound CoSOD performance.

✓ Oracle masks from **SAM** already achieve **SOTA upper-bound performance**.

2. Methodology

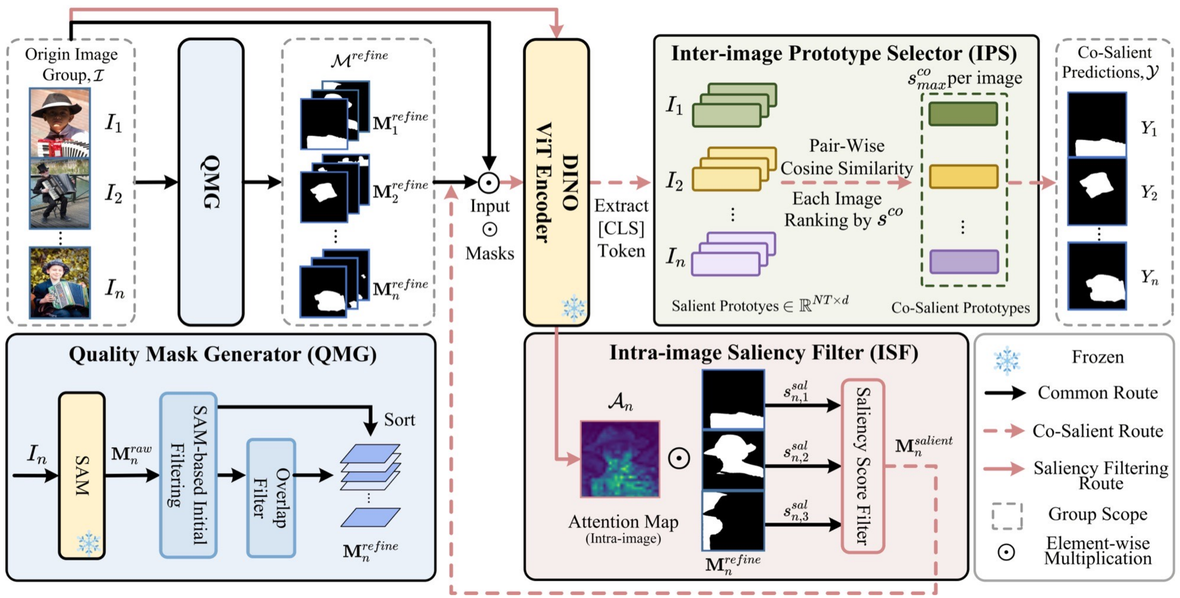

Fig. 2. Overview of the TF-SSD pipeline: QMG filters SAM masks, ISF applies DINO saliency, IPS selects cross-image consistent masks.

QMG removes noisy/redundant SAM masks. **ISF** uses DINO attention to keep salient masks per image. **IPS** selects cross-image consistent masks via prototype matching.

1) Quality Mask Generator (QMG)

- Area filter: keep masks with $r^{\text{area}} \in [r_{\min}, r_{\max}]$
- Overlap removal (NMS-style): remove m_j if $\rho_{i\to j} > \theta_{\text{ov}}$
- Quality score: $q = \alpha \cdot \text{IoU}_{\text{SAM}} + \beta \cdot s^{\text{size}}$
- Select top-K masks per image

2) Intra-image Saliency Filter (ISF)

- DINO [CLS] attention map $\mathcal{A}_n$
- Score: $s^{\text{sal}} = \frac{1}{|m|}\Sigma \mathcal{A}_n \odot m$
- Keep top-T=6 per image

3) Inter-image Prototype Selector (IPS)

- Prototype $p_{n,t} = \mathcal{F}(I_n \odot m_{n,t})$
- Pair-wise cosine similarity
- Max-similarity cross-image matching

Fig. 3. IPS selects masks with highest cross-image prototype similarity.

3. Experiment

Performance on CoCA:

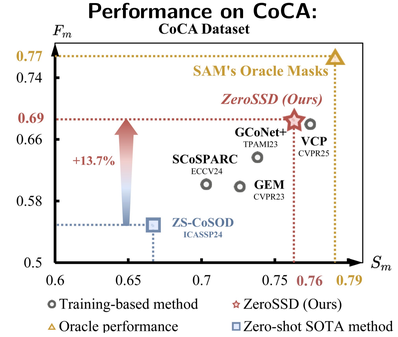

Fig. 4. F_m–S_m scatter plot on CoCA. **TF-SSD** (star) surpasses all prior methods.

Comparison on Multiple Benchmarks:

Method	T	CoCA $F_\beta\uparrow$	CoCA $S_\alpha\uparrow$	CoSal2015 $F_\beta\uparrow$	CoSal2015 $S_\alpha\uparrow$	CoSOD3k $F_\beta\uparrow$	CoSOD3k $S_\alpha\uparrow$
DCFM CVPR22	S	0.598	0.710	0.856	0.838	0.805	0.810
GCoNet+ TPAMI23	S	0.637	0.738	0.891	0.881	0.834	0.843
CONDA ECCV24	S	0.685	0.763	0.908	0.900	0.853	0.862
VCP CVPR25	S	0.680	0.774	0.920	0.911	0.868	0.874
ZS-CoSOD ICASSP24	TF	0.549	0.667	0.799	0.785	0.691	0.723
TF-SSD	TF	**0.686**	**0.763**	**0.899**	**0.890**	**0.860**	**0.861**

Tab. 1. Quantitative comparison on CoCA, CoSal2015, and CoSOD3k. Best results **bold**. **S**: Supervised. **TF**: Training-free.

Qualitative Comparison:

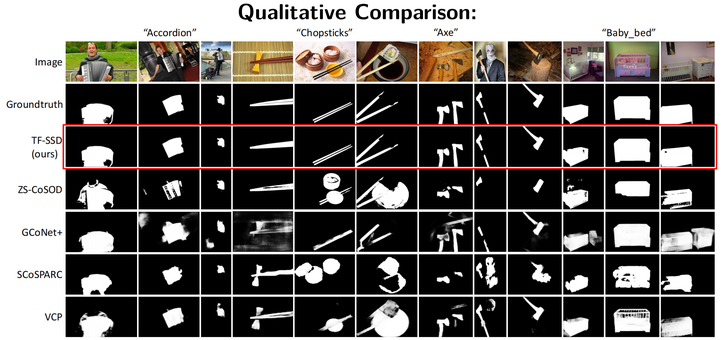

Fig. 5. Qualitative comparison.

★ TF-SSD **outperforms all methods**, surpassing **ZS-CoSOD** by **+13.7%** in F_m and **+9.6%** in S_m.

TF-SSD matches or exceeds supervised methods **without any training**, demonstrating the power of foundation model synergy.

TF-SSD (red box) yields **more complete & accurate masks** — thin ("Chopsticks"), occluded ("Axe"), cluttered ("Baby_bed").

4. Conclusion

- **TF-SSD** is a **training-free** CoSOD framework built on **SAM** × **DINO** synergy.
- **QMG** → **ISF** → **IPS**: a clean three-stage pipeline requiring *no labeled data*.
- **Achieves state-of-the-art results** against both supervised and training-free methods on CoCA, CoSal2015 and CoSOD3k.
- Demonstrates the strong potential of VFMs for CoSOD in a training-free manner.
- **Limitation:** performance depends on the quality of SAM masks and DINO features.
- **Future work:** extend to video CoSOD and integrate stronger VFMs.

Training-free · No annotation · VFM powered

* Equal contribution † Corresponding author