# Advancing Complex Video Object Segmentation via Progressive Concept Construction

Zhixiong Zhang*, Shuangrui Ding*, Xiaoyi Dong†, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, Jiaqi Wang†

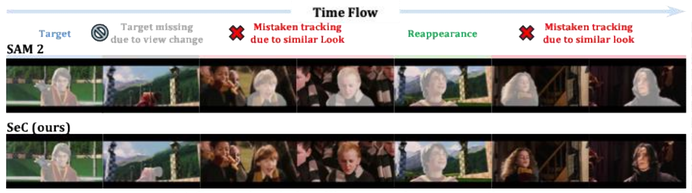
上海创智学院 Shanghai Innovation Institute

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

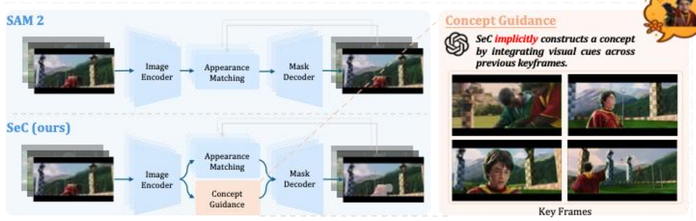
ICLR International Conference On Learning Representations

## Motivation

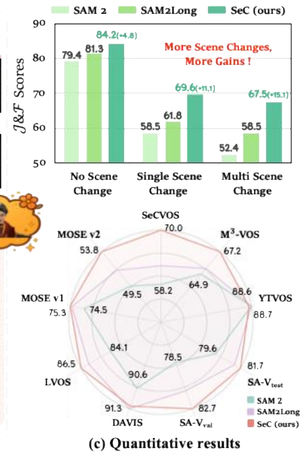
- SAM 2 struggles when facing appearance-similar distractors or abrupt scene transitions.
- If the model understands the object's role (e.g., player or audience) as a concept, can it better resolve SAM 2's limitations in tracking consistency?




(a) Comparison in handling scene changes over time.




(b) Schematic comparison between our SeC and SAM 2



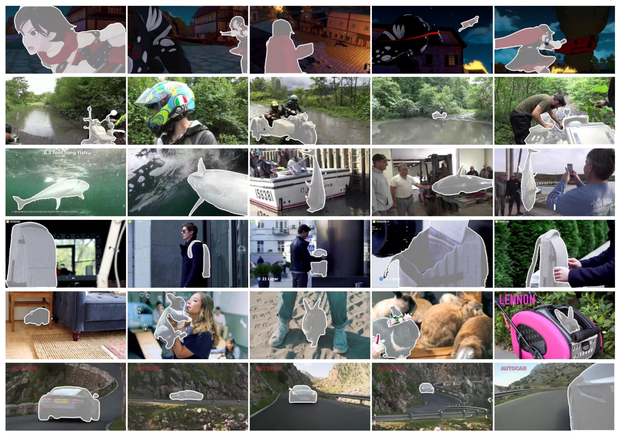

(c) Quantitative results

## SeCVOS

**A Benchmark for Semantically Complex VOS**

- 160 manually annotated multi-shot video clips;
- Frequent target reappearance and high temporal discontinuity;
- Designed to evaluates performance under high-level semantic reasoning;

**Time Flow**



## Experiment Results

- SeC consistently performs better across benchmarks, with more scene changes leading to greater improvements.

| Method | No Scene Change | | | Single Scene Change | | | Multi Scene Change | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
| Xmem (Cheng & Schwing, 2022) | 71.9 | 72.0 | 71.8 | 47.0 | 47.9 | 46.2 | 41.9 | 42.4 | 41.4 | 48.4 |
| DEVA (Cheng et al., 2023) | 71.6 | 71.6 | 71.5 | 48.5 | 48.4 | 48.6 | 46.4 | 46.0 | 46.8 | 49.7 |
| Cutie-base (Cheng et al., 2024) | 72.5 | 72.2 | 72.8 | 53.0 | 52.9 | 53.2 | 48.3 | 47.8 | 48.9 | 52.7 |
| SAM2.1 (Ravi et al., 2025) | 79.4 | 79.1 | 79.7 | 58.5 | 58.2 | 58.8 | 52.4 | 52.1 | 52.6 | 58.2 |
| SAMURAI (Yang et al., 2024) | 81.8 | 81.6 | 81.9 | 60.6 | 60.6 | 60.7 | 59.3 | 58.9 | 59.7 | 62.2 |
| SAM2.1Long (Ding et al., 2025d) | 81.3 | 81.0 | 81.6 | 61.8 | 61.6 | 62.0 | 58.5 | 58.1 | 58.9 | 62.3 |
| **SeC (Ours)** | **84.2**$_{+4.8}$ | **83.8** | **84.5** | **69.6**$_{+11.1}$ | **69.5** | **69.7** | **67.5**$_{+15.1}$ | **67.0** | **68.0** | **70.0**$_{+11.8}$ |

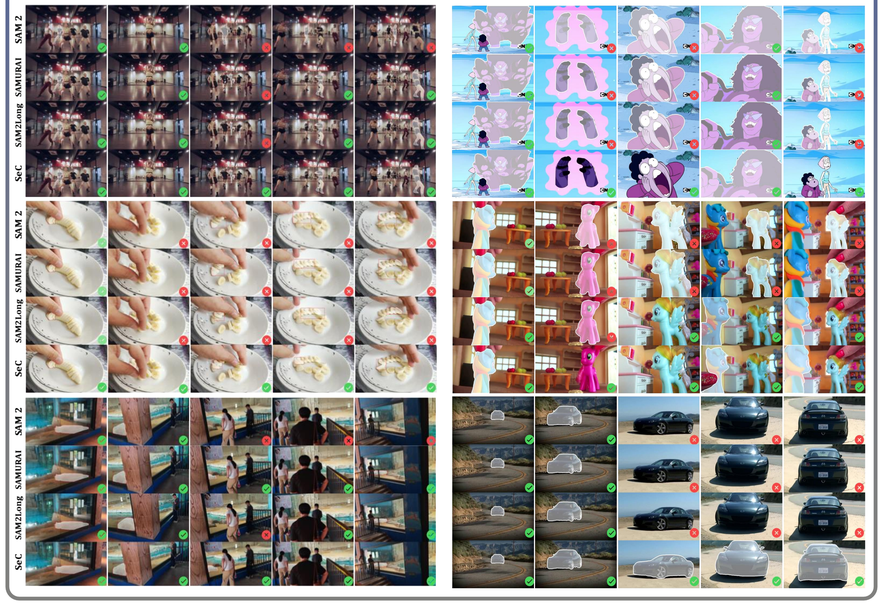
| Pixel-level Association | Concept Guidance | SA-V $\mathcal{J}\&\mathcal{F}$ | SeCVOS $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|
| ✗ | ✗ | 78.6 | 58.2 |
| ✓ | ✗ | 82.4 | 62.2 |
| ✓ | ✓ | **82.7** | **70.0** |

| LVLM Size | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| 1B | 68.4 | 68.2 | 68.7 |
| 2B | 69.5 | 69.3 | 69.8 |
| 4B | 70.0 | 69.7 | 70.2 |
| 8B | 70.3 | 70.1 | 70.7 |

| Method | $\mathcal{J}\&\mathcal{F}$ | | | | | $\mathcal{G}$ | $\mathcal{J}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|---|---|---|---|---|
| | SA-V val | SA-V test | LVOS v2 val | MOSE v1 val | DAVIS 2017 val | YTVOS 2019 val | $M^3$-VOS core | MOSE v2 val |
| STCN (Cheng et al., 2021) | 61.0 | 62.5 | 60.6 | 52.5 | 85.4 | 82.7 | - | 29.7 |
| SwinB-AOT (Yang et al., 2021) | 51.1 | 50.3 | - | 59.4 | 85.4 | 84.5 | - | 30.2 |
| SwinB-DeAOT (Yang & Yang, 2022) | 61.4 | 61.8 | 63.9 | 59.9 | 86.2 | 86.1 | 62.3 | 32.6 |
| RDE (Li et al., 2022) | 51.8 | 53.9 | 62.2 | 46.8 | 84.2 | 81.9 | - | 32.0 |
| XMem (Cheng & Schwing, 2022) | 60.1 | 62.3 | 64.5 | 59.6 | 86.0 | 85.6 | 60.6 | 36.3 |
| SimVOS-B (Wu et al., 2023) | 44.2 | 44.1 | - | - | 88.0 | 84.2 | - | - |
| DEVA (Cheng et al., 2023) | 55.4 | 56.2 | - | 66.0 | 87.0 | 85.4 | - | 38.3 |
| ISVOS (Wang et al., 2023) | - | - | - | - | 88.2 | 86.3 | - | - |
| TarVIS (Athar et al., 2023) | - | - | - | - | 85.2 | - | - | - |
| UNINEXT (Yan et al., 2023) | - | - | - | - | 81.8 | 78.6 | - | - |
| UniVS (Li et al., 2024) | - | - | - | - | 76.2 | 71.5 | - | - |
| JointFormer (Zhang et al., 2025b) | - | - | - | - | 90.1 | 87.4 | - | 37.7 |
| Cutie-base (Cheng et al., 2024) | 60.7 | 62.7 | - | 69.9 | 87.9 | 87.0 | 64.6 | 42.8 |
| Cutie-base+ (Cheng et al., 2024) | 61.3 | 62.8 | - | 71.7 | 88.1 | 87.5 | - | - |
| SAM 2.1 (Ravi et al., 2025) | 78.6 | 79.6 | 84.1 | 74.5 | 90.6 | **88.7** | 64.9 | 49.5 |
| SAMURAI (Yang et al., 2024) | 79.8 | 80.0 | 84.2 | 72.6 | 89.9 | 88.3 | - | 51.1 |
| SAM2.1Long (Ding et al., 2025d) | 81.1 | 81.2 | 85.9 | 75.2 | **91.4** | **88.7** | 65.5 | 51.5 |
| **SeC (Ours)** | **82.7** | **81.7** | **86.5** | **75.3** | 91.3 | 88.6 | **67.2** | **53.8** |

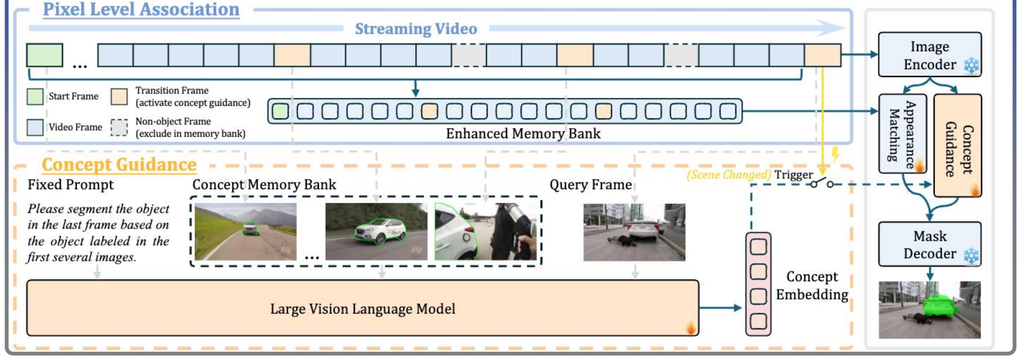
| #Concept Tokens | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| 1 | 70.0 | 69.7 | 70.2 |
| 2 | 70.0 | 69.7 | 70.3 |
| 4 | 69.9 | 69.6 | 70.2 |

| Scene Detector | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| ABS DIFF | 69.0 | 68.8 | 69.3 |
| ORB | 68.7 | 68.5 | 69.0 |
| SSIM | 69.3 | 69.1 | 69.6 |
| FLOW | 69.5 | 69.3 | 69.7 |
| HSV (ours) | 70.0 | 69.7 | 70.2 |




## Method: Progressive Concept Construction

- SeC leverage the capability of LLMs to introduce concept-level features.
- Concept Guidance only activate when the scene is changed and concept bank is updated dynamically.




## Discussion

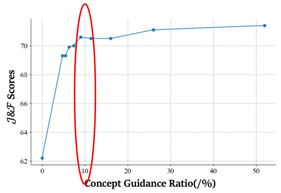
**Does SeC progressively construct concept-level representation?**

- SeC learns comprehensive concepts on the fly.
- Offline setting even makes the performance stronger.

| Concept construction | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| None | 62.2 | 61.8 | 62.6 |
| Online | 70.0 | 69.7 | 70.2 |
| Offline | 71.8 | 71.5 | 72.1 |

**Does SeC require frequent concept guidance?**

- Activating less than **10%** of frames is sufficient;
- Good trade-off between accuracy and efficiency.




| Benchmark | Method | $\mathcal{J}\&\mathcal{F}$ | Con. Guid. Ratio (%) | Throughput ($s^{-1}$) |
|---|---|---|---|---|
| SeCVOS | SeC | 70.0 | 7.4 | 14.8 |
| | SAM 2 | 58.2 | N/A | 22.0 |
| SA-V | SeC | 82.7 | 1.0 | 18.1 |
| | SAM 2 | 78.6 | N/A | 22.0 |