# Strike the Balance: On-the-Fly *Uncertainty* based *User Interactions* for *Long-Term Video Object Segmentation*

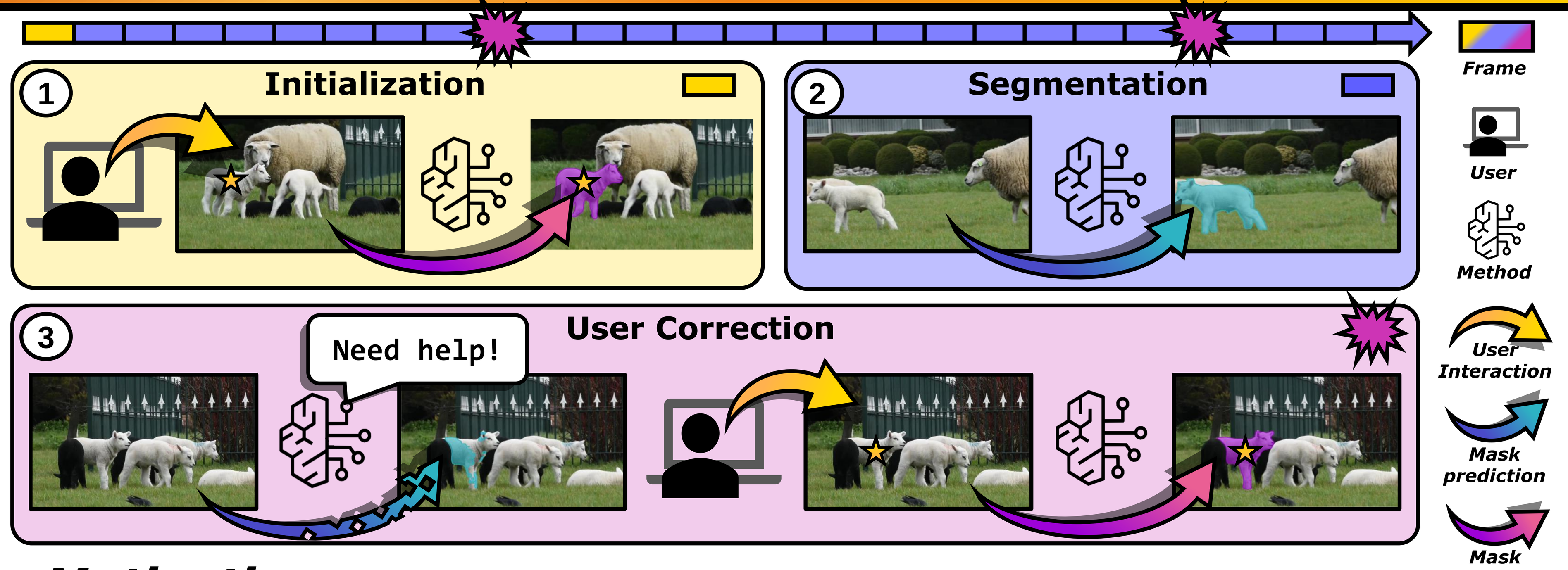

**Stéphane Vujasinović**[1], Stefan Becker[1], Sebastian Bullinger[1], Norbert Scherer-Negenborn[1], Michael Arens[1] and Rainer Stiefelhagen[2]

[1] Fraunhofer IOSB (Ettlingen, Germany)
[2] Karlsruhe Institute of Technology (Karlsruhe, Germany)


Scan for Code

Fraunhofer IOSB | KIT Karlsruher Institut für Technologie | ACCV 2024

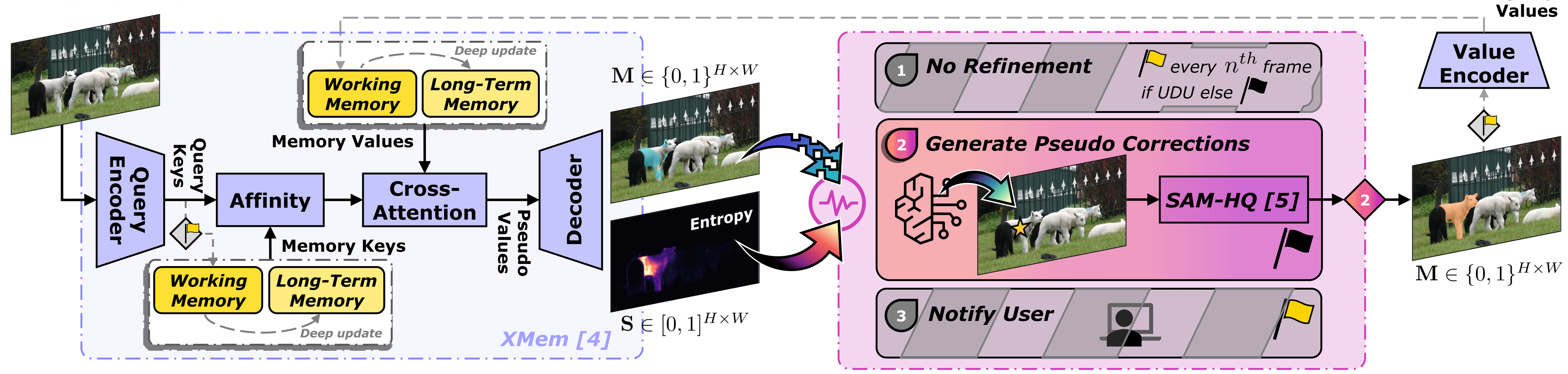
## TL;DR




- **Motivations**
  - Maximize tracking for long-term VOS
  - Minimize human oversight (only at delicate events)
  - Allow user corrections on-the-fly with user-clicks
- **Contributions: *ziVOS* and *Lazy-XMem***
  - New Task: Lazy Video Object Segmentation
  - On-the-fly assessment of tracking accuracy
  - Generate pseudo- and support user-corrections on-the-fly
- **Results**
  - Increase accuracy (by 11% $\mathcal{J}\&\mathcal{F}$) and robustness (by 10% $R@0.1$)
  - Interact with only 1% of the dataset to improve robustness

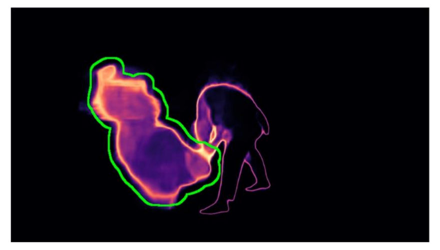
## Methodology – Lazy-XMem

### Overview

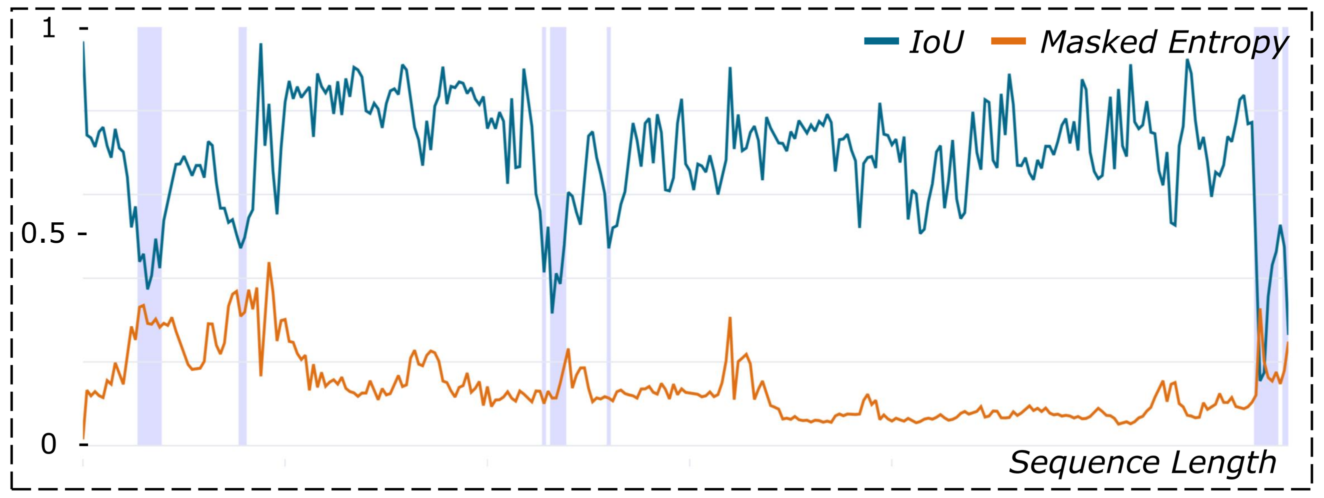



### 1. Uncertainty Estimation

- Pixel-level uncertainty [1]

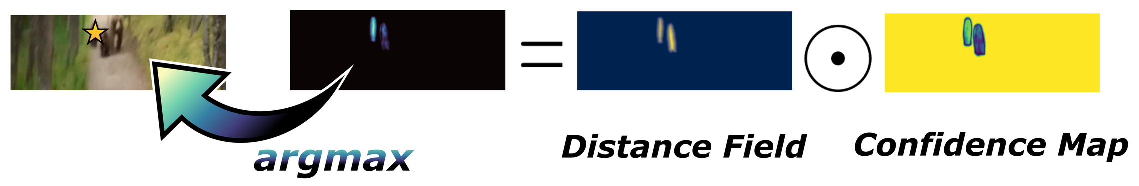


- IoU vs Uncertainty (Correlation? Yes!)




### 2. Mask Refinement

- Sam-HQ [5]
- Issue corrections
- Pseudo-corrections on-the-fly




### 3. Memory Update

ON OFF

- Uncertainty driven update (UDU)
- Interaction driven update (IDU)

## Metrics

*New metrics for ziVOS

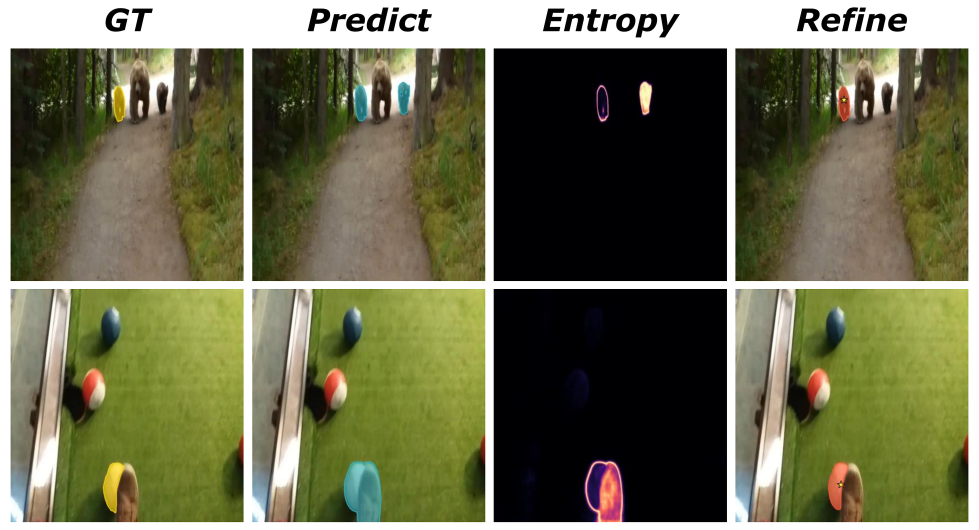
- **Performance**
  - Accuracy ($\mathcal{J}\&\mathcal{F}$)
  - Robustness ($R@\tau_{\mathrm{IoU}}$)
- **User-Workload**
  - Number of corrections (NoC)
  - Interaction density index (IDI)
  - Average correction interval (ACI)

$$R@\tau_{\mathrm{IoU}} = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{|\mathcal{F}_o|}\sum_{f\in\mathcal{F}_o}\mathbb{1}_{[\mathrm{IoU}(\mathbf{M}_f^o,\mathbf{GT}_f^o)\geq\tau_{IoU}]} \qquad \mathrm{ACI} = \sum_{o\in\mathcal{O}}\frac{1}{|\mathcal{F}_o|}\sum_{i=1}^{|\mathcal{F}_o|}\sum_{j=1}^{i} n_j$$

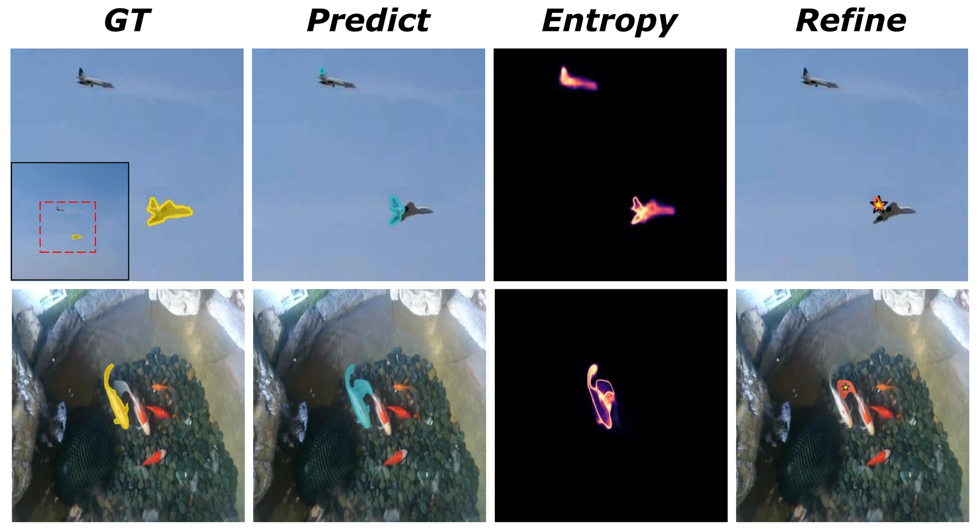
## Qualitative Results

### Pseudo-Corrections

Successes

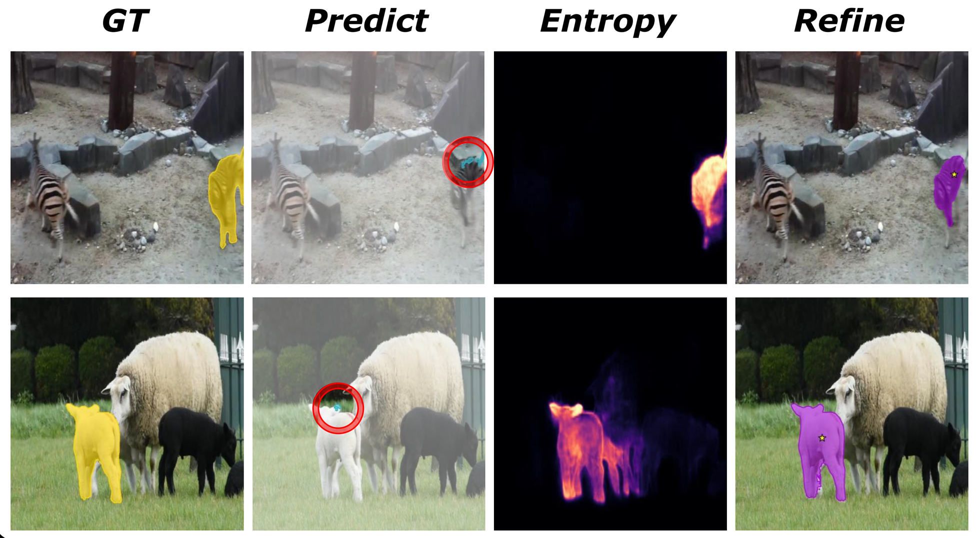



Failures

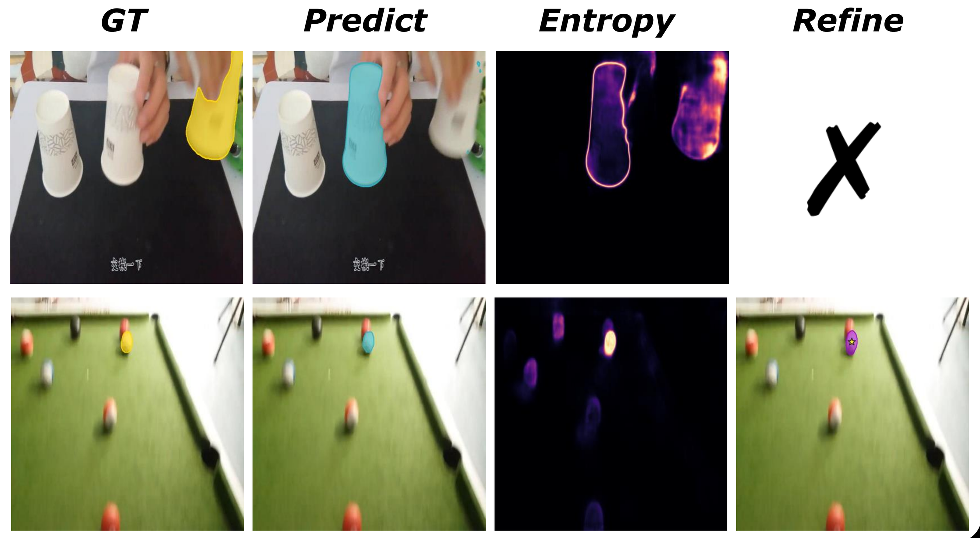



### User-Corrections

Successes




Failures




## Quantitative Results – ziVOS

### Benchmark

| Method | $\mathcal{J}\&\mathcal{F}$ | Robustness $R@0.1$ | $R@0.25$ | $R@0.5$ | User-Workload ACI | NoC | IDI |
|---|---|---|---|---|---|---|---|
| *sVOS Methods* | | | | | | | |
| QDMN [6] (ECCV 2022) | 44.2 | 47.8 | 45.5 | 36.2 | - | - | - |
| XMem [4] (ECCV 2022) | 52.8 | 57.0 | 55.0 | 49.0 | - | - | - |
| DEVA [7] (ICCV 2023) | 55.1 | **63.6** | **59.3** | 52.4 | - | - | - |
| Cutie-base [8] (CVPR 2024) | 57.0 | 59.2 | 57.8 | 52.4 | - | - | - |
| Cutie-small [8] (CVPR 2024) | **57.6** | 58.6 | 57.0 | **52.5** | - | - | - |
| Lazy-XMem† (ours) | 56.4 | 58.8 | 56.8 | 50.6 | - | - | - |
| *ziVOS Methods* | | | | | | | |
| Rand-Lazy-XMem | 61.3 | 67.9 | 65.8 | 59.3 | 5.17 | 335 | 17.9 |
| Lazy-QDMN | 52.7 | 58.2 | 52.0 | 42.9 | 5.64 | 360 | 16.7 |
| Lazy-XMem (ours) | **64.3** | **70.2** | **67.8** | **62.3** | 5.02 | **325** | 18.4 |

† No User Corrections

### Ablations

Results from LVOS [3]

| Configuration UDU | *Pseudo* Corr. | *Pseudo* IDU | *User* Corr. | *User* IDU | $\mathcal{J}\&\mathcal{F}$ | Robustness $R@0.1$ | $R@0.25$ | $R@0.5$ | User-Workload ACI | NoC | IDI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | 52.8 | 57.0 | 55.0 | 49.0 | - | - | - |
| ✓ | - | - | - | - | 54.7 | 56.3 | 54.5 | 50.0 | - | - | - |
| ✓ | ✓ | - | - | - | 56.4 | 58.8 | 56.8 | 50.6 | - | - | - |
| ✓ | ✓ | ✓ | - | - | 53.1 | 57.0 | 55.1 | 49.6 | - | - | - |
| ✓ | - | - | ✓ | - | 55.6 | 58.2 | 56.4 | 51.8 | 7.80 | 507 | 12.6 |
| ✓ | - | - | ✓ | ✓ | 62.9 | 67.8 | 66.2 | 60.9 | 5.05 | 327 | 18.3 |
| ✓ | ✓ | - | ✓ | ✓ | **64.3** | **70.2** | 67.8 | **62.3** | 5.02 | **325** | 18.4 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **64.3** | 70.1 | **68.2** | 62.1 | 5.91 | 352 | 17.3 |


[1] Shannon, C.E. A mathematical theory of communication. The Bell system technical journal, 1948.
[2] Pont-Tuset, J. *et al.* The 2017 DAVIS challenge on video object segmentation. arXiv, 2017.
[3] Hong, L., *et al.* LVOS: A benchmark for long-term video object segmentation. ICCV, 2023.
[4] Cheng, H.K., Schwing, A.G. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. ECCV, 2022.
[5] Ke, L., *et al.* Segment anything in high quality. NeurIPS, 2023
[6] Liu, Y., *et al.* Learning quality aware dynamic memory for video object segmentation. ECCV, 2022.
[7] Cheng, H.K., *et al.* Tracking anything with decoupled video segmentation. ICCV, 2023.
[8] Cheng, H.K., *et al.* Putting the object back into video object segmentation. CVPR, 2024.



Contact: stephane.vujasinovic@iosb.fraunhofer.de