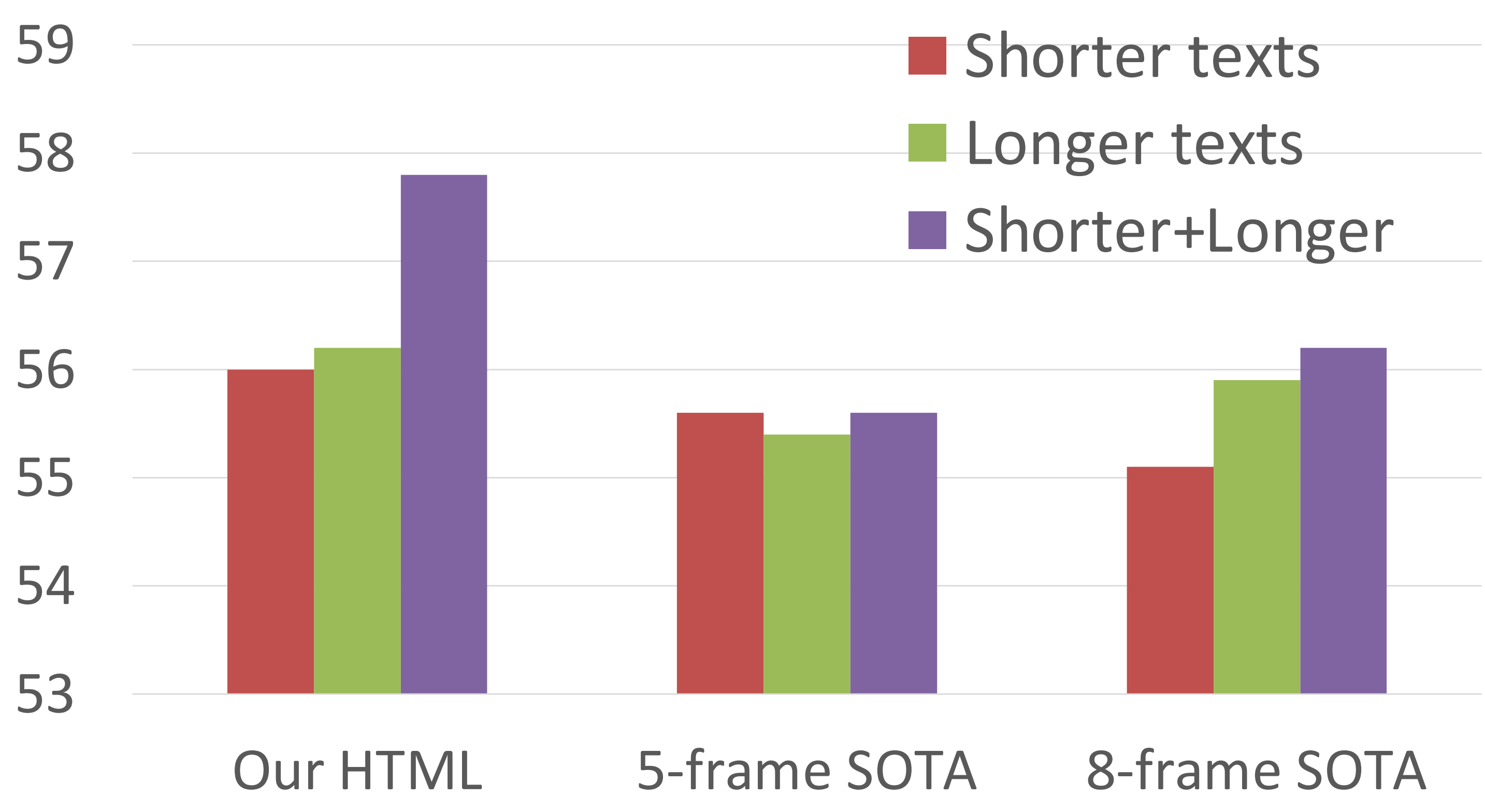
HTML: Hybrid Temporal Multimodal Learning Framework for Referring Video Segmentation

Mingfei Han[1,4], Yali Wang[2,6]✉, Zhihui Li[3], Lina Yao[4], Xiaojun Chang[1,5], Yu Qiao[6,2]

[1] ReLER Lab, AAII, UTS [2] SIAT, CAS [3] Shandong Artificial Intelligence [4] Data61, CSIRO [5] MBZUAI [6] Shanghai AI Lab

Highlights:

- ➢ Our method with ResNet-50 achieves **57.8** in L&F, surpassing the recent SOTA with ResNet-101.
- ➢ Our HTML boosts the baseline model **without additional** modules and computations during inference.
- ➢ Our HTML can significantly benefit from **diversified** text descriptions.



Referring Video Object Segmentation

Given:

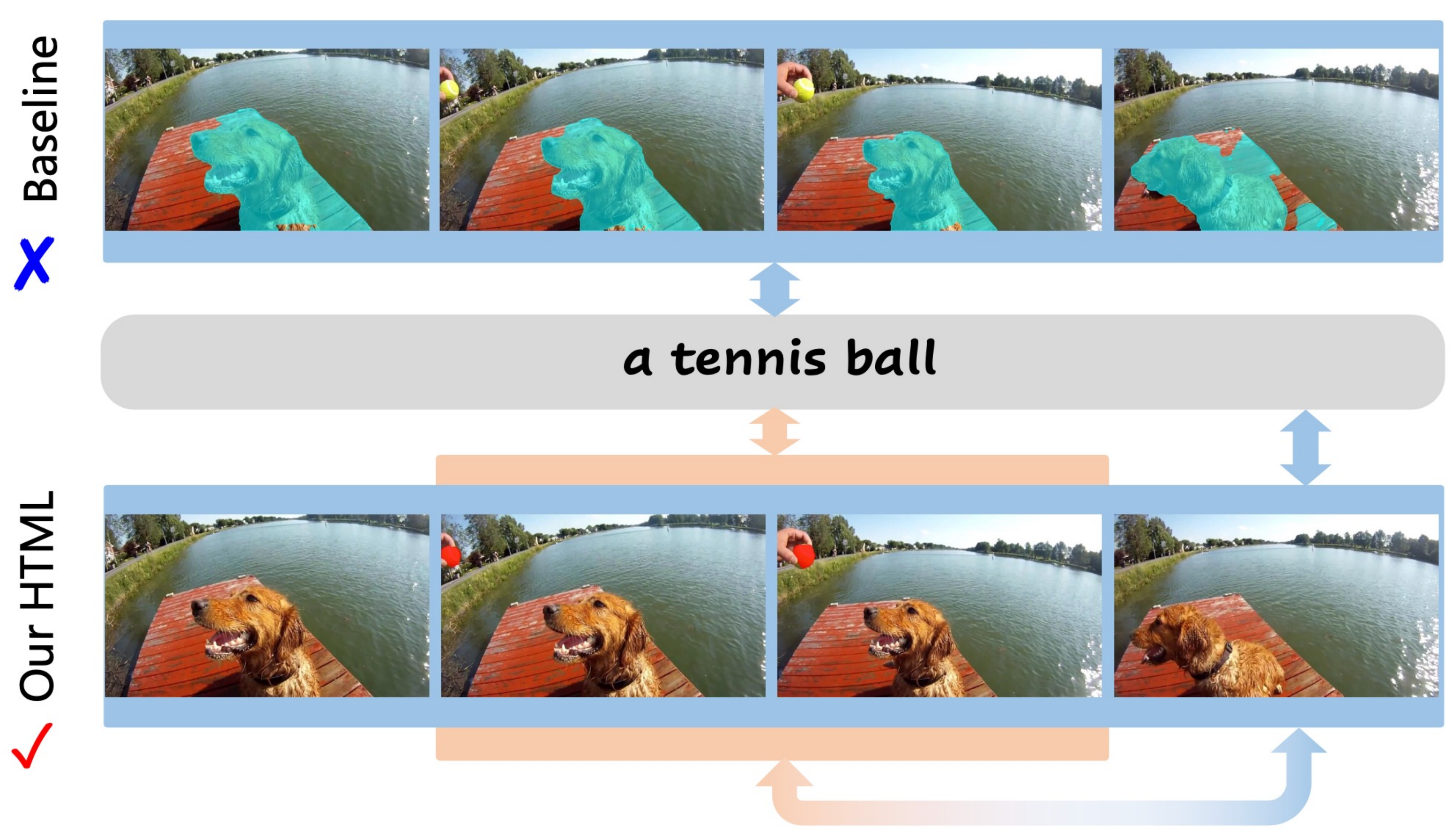
- A text sentence of the object in the video, describing the motion, appearance and positions.
- A video containing the interested object.

Output:

- The mask of the object in each frame of the video.

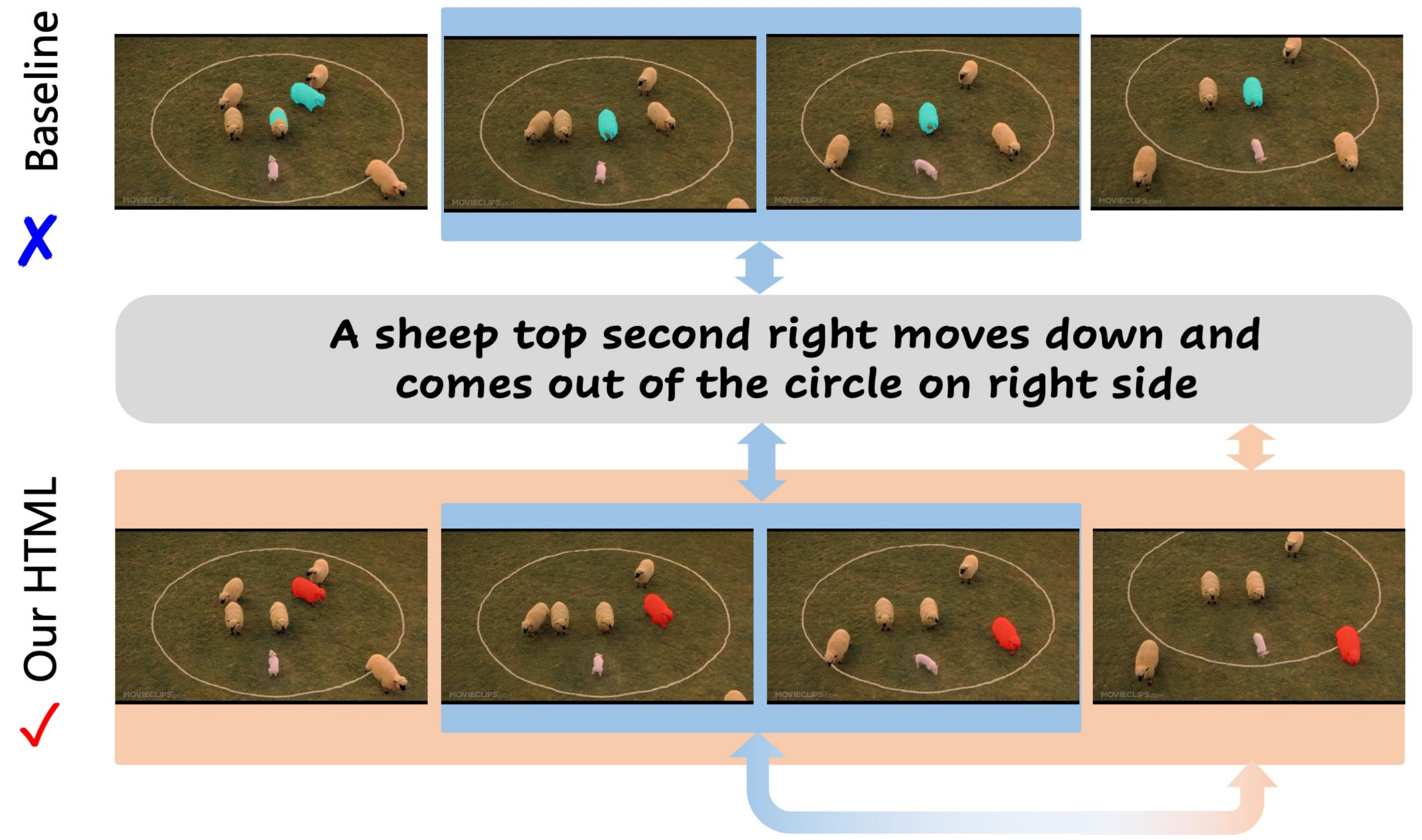
On Academic Job Market

Motivation:



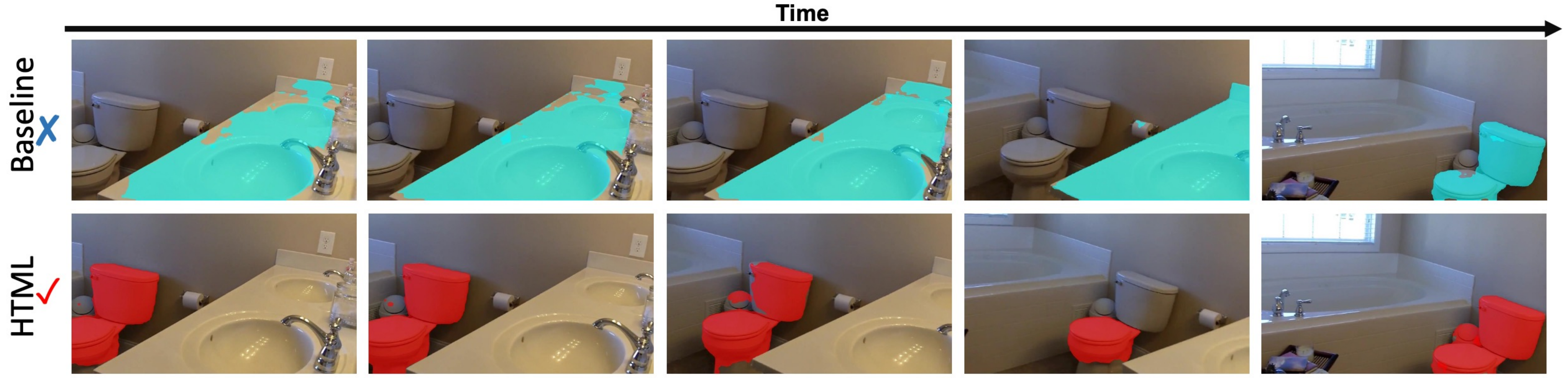
(a) Simple description

- ❖ An object can be described with language descriptions in different lengths.
- ❖ Texts in different complexities relate to different temporal lengths.


Baseline ✗
A sheep top second right moves down and comes out of the circle on right side
Our HTML ✓

(b) Complex description

Visualization:



(a) The white toilet is behind the two sinks in the bathroom

- ❖ Baseline simply uses a single temporal scale.
- ❖ With hybrid temporal scales, our HTML can discover the object semantics.

Contact us: Please visit https://mingfei.info/HTML or scan the QR Code.

ICCV23 PARIS