**Motivation of This Work**

- existing multi-stage pipeline: localize and recognize in each frame, track for text streams, then post-processing. Two problems:
  - excessive computation cost from repetitive recognition
  - unstable recognition results due to low-quality text

**Main Contributions**

- an unified two-stage framework YORO consisting of a spatial-temporal detector and a text recommender for fast video text spotting.
- a novel text recommender for selecting the highest-quality text from streams, then only recognizing the selected text regions once.

**Text Recommender**

- a spatial-temporal detector for robustly recall more text by referring to temporal relationship among different frames.
- self-attention based aggregation

**Key Component**

- mechanism of quality scoring network

**Experiments & Ablation**

- **ablation:**
  - performance and speed comparison with other frame selection methods
  - effectiveness of each module

**Proposed Dataset (LSVTD)**

- existing video scene text datasets: limited scale and scenes, which may restrain research of video scene text spotting.
- our collected dataset:
  - 22 indoor/outdoor real-world scenarios (100 videos)