"Previously On ..." From Recaps to Story Summarization

Aditya Kumar Singh, Dhruv Srivastava, Makarand Tapaswi

Katha AI

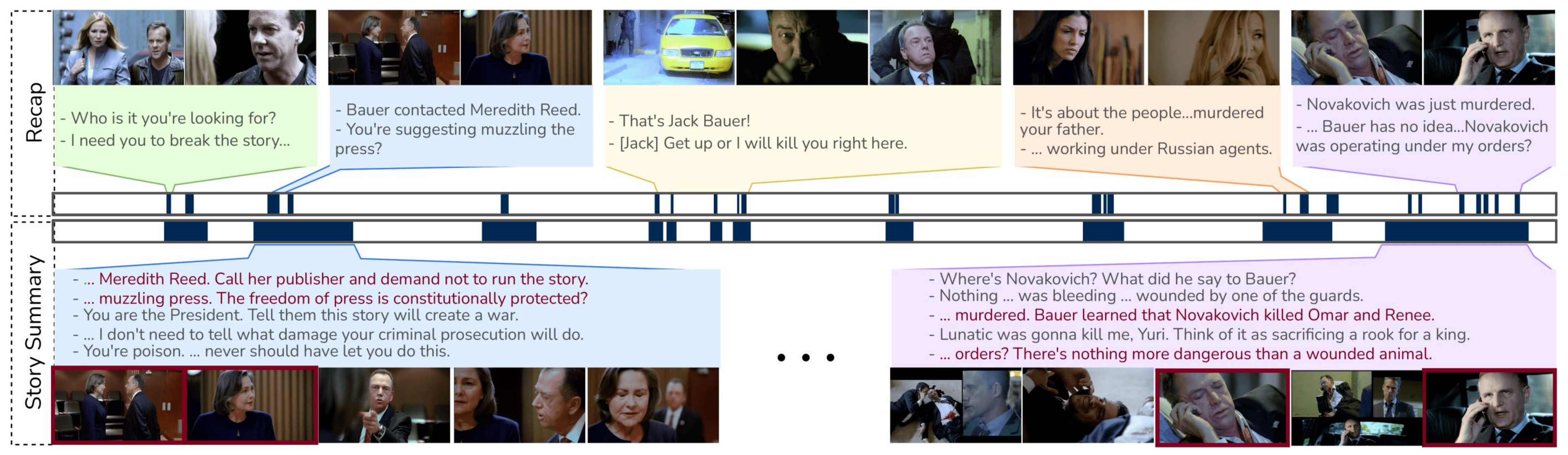
International Institute of Information Technology Hyderabad

CVPR Seattle, WA June 17-21, 2024

https://katha-ai.github.io/projects/recap-story-summ/

Introduction

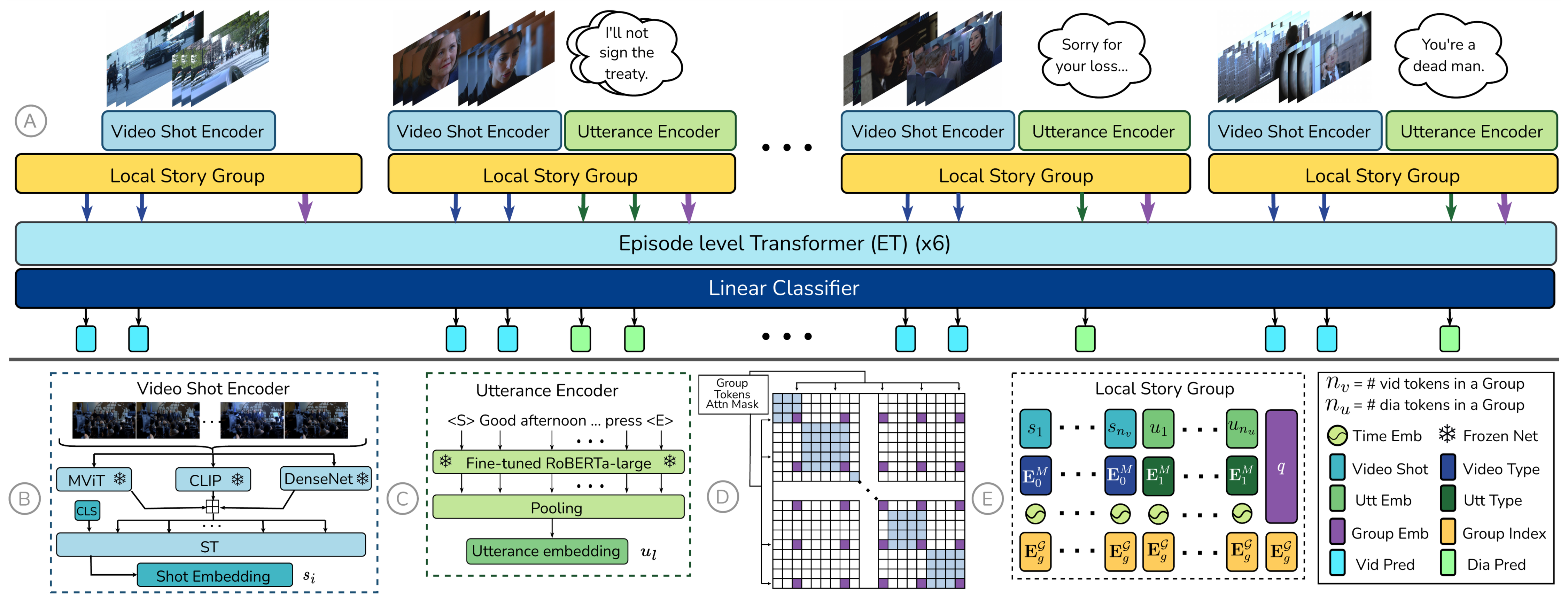
➢ **Goal**: Our aim is to extract a multimodal *"story"* summary (video and text) from a given episode of a TV Show, which typically lasts around 40 minutes.



➢ **Motivation**

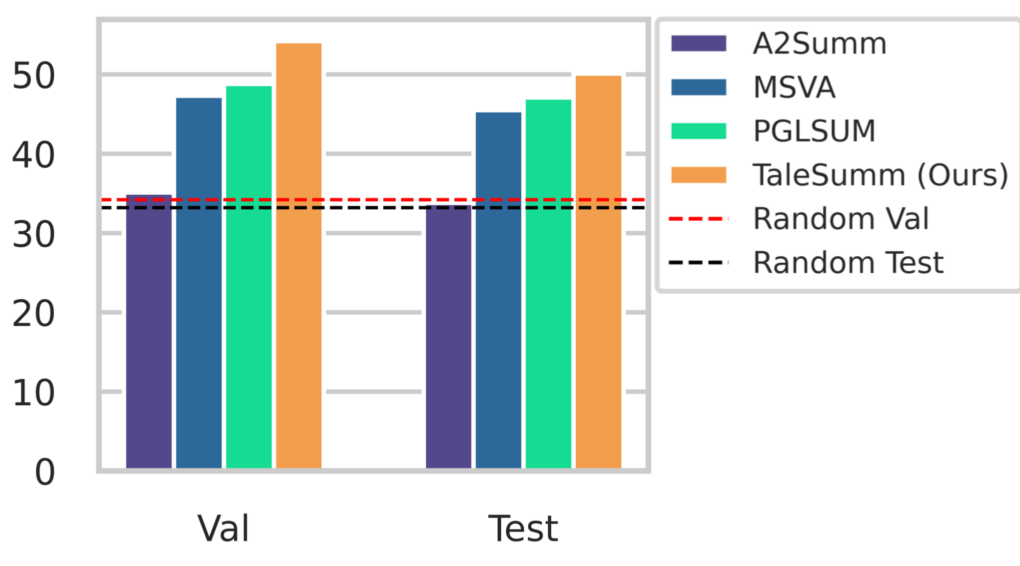
1. Typical video summarization focus on generating keyframes, skims, video storyboards or synopses. While in space of text, focus is on text-summary.
2. There are multimodal (input) to unimodal / multimodal (output) approaches too, but we significantly differ in the type of video (stories *vs.* creative/ documentary).
3. Leveraging ***recap*** (as our supervising signal), we tried to capture the overall story-arc of an episode (~**40** mins) where temporal video/text signal may not align semantically, making this task even more challenging.

TaleSumm: Our Approach for Story Summarization

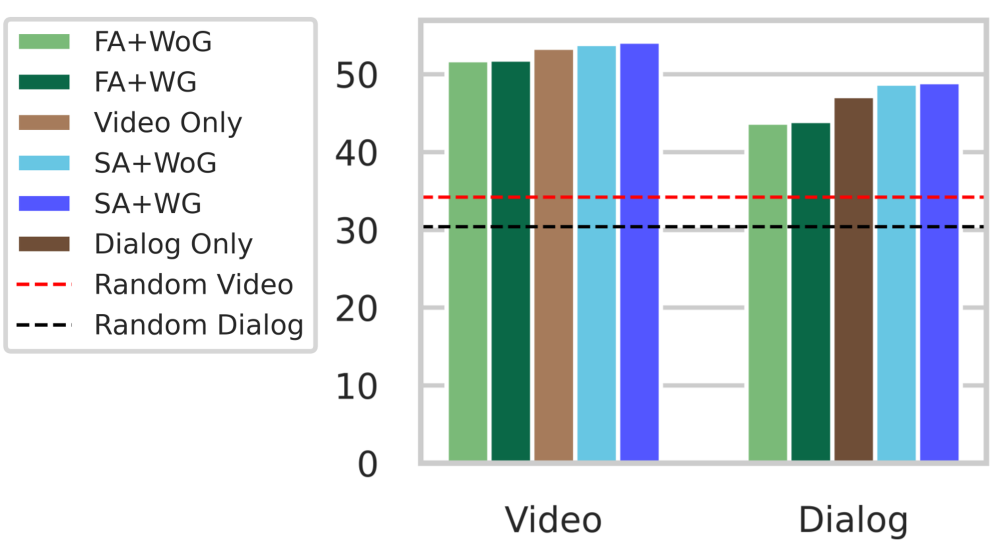


1. **Level 1: Shot/Dialog Representation**
   a. Prep **Shot/Dialog encodings**: Raw video shots and dialogs pass through **B** and **C**, respectively.
   b. Form Story Groups from temporally arranged **V**/**D** tokens with an appended group token.
2. **Level 2: Episode-Level Interactions**
   a. Capture **video/dialog tokens interactions** within their corresponding story groups.
   b. Pass the aggregated info inside each SG across every SGs via **GT** with ***special*** attention mask.
   c. A shared linear classifier at the end for video-shot/dialog to predict their importance.
   d. Our approach TaleSumm is trained in an end-to-end fashion with *BCE* loss.

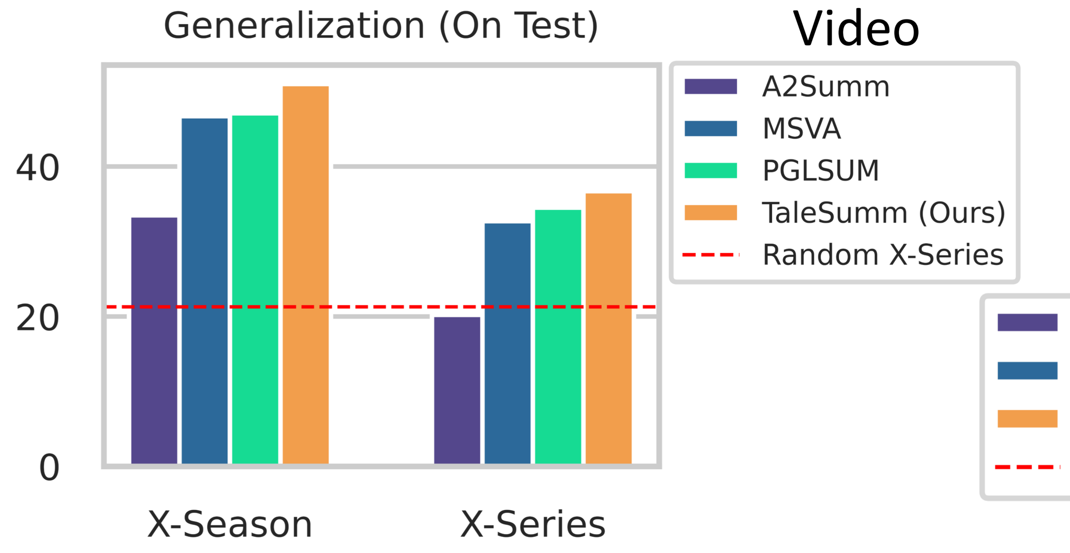
Experiments (Metric: AP)

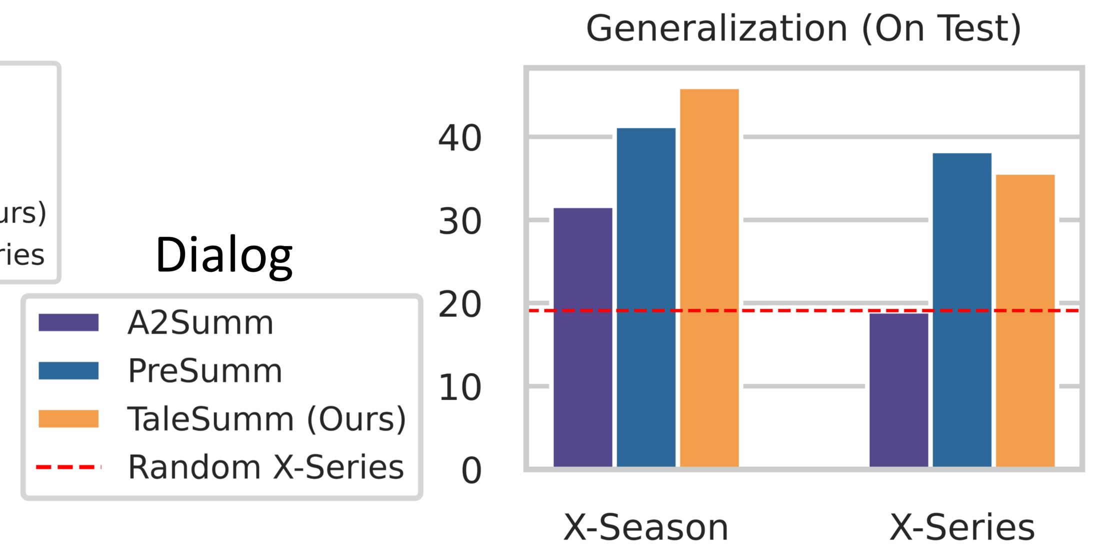


TaleSumm outperforms SoTA models adapted for our task.



Special Attention with Group Token setting is **BEST**.





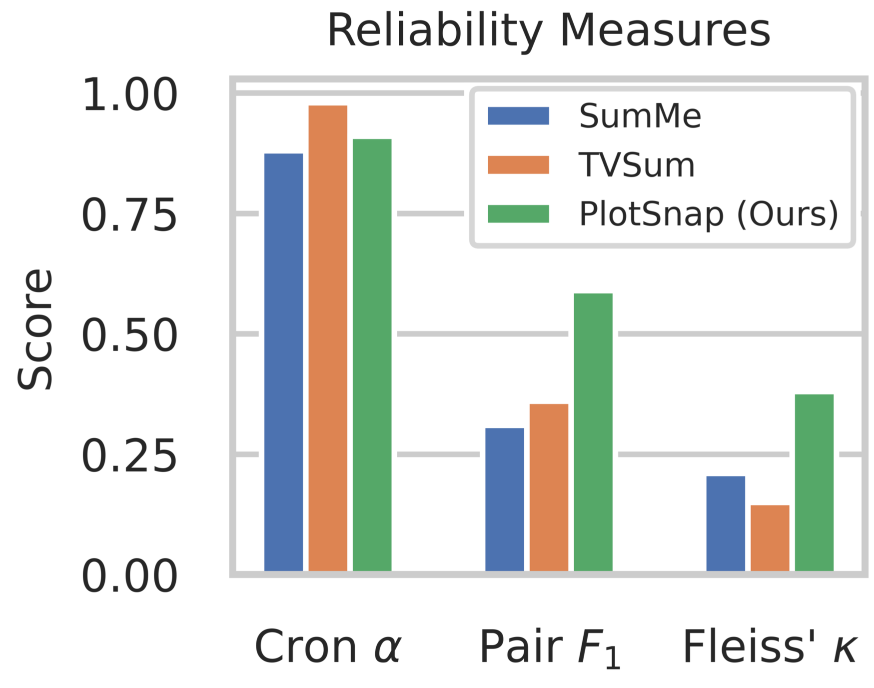
TaleSumm generalizes better on X-Season/Series (an entire season/series in test)

PlotSnap: Our Multimodal Dataset

2 famous crime-thriller TV Shows: (i) **24**, (ii) **Prison Break**

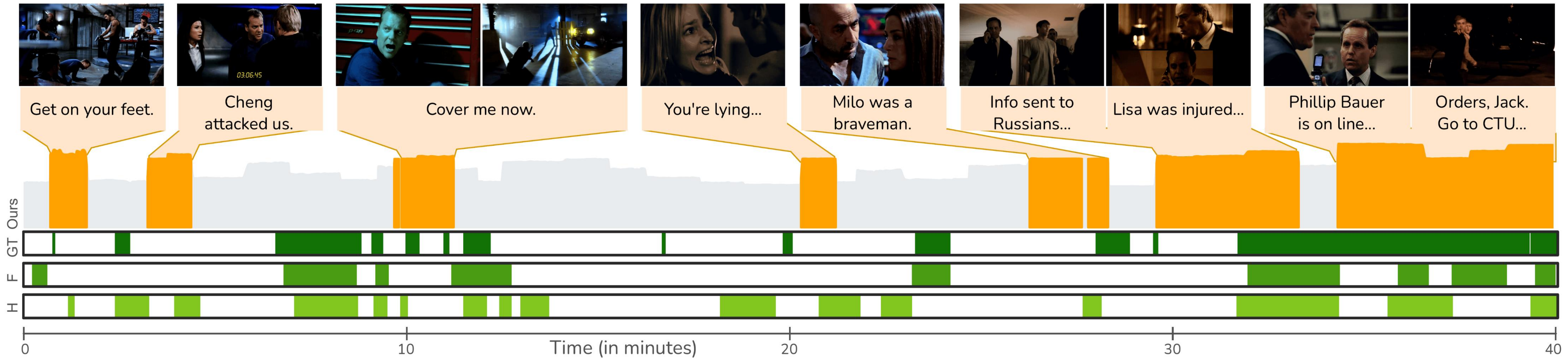
- **Why action thrillers?** Challenging than rom-/sit-coms and captivating plot-lines (story-arc).
- **Key Idea:** Use **Recaps** from next episode to extract story-summaries from the current one.
- **Label Extraction:** (i) shot-matching, (ii) smoothing
- **Data Stats:** Altogether, **10** Seasons with **205** episodes.

TV Series	24	Prison Break
# of Seasons	8	2
# of Episodes	172	33
Dataset duration (hours)	125.9	24.0
Avg episode duration (s)	2635 ± 72	2615 ± 39
Avg # of shots per episode	825 ± 101	999 ± 117
Avg duration of shots (s)	3.2 ± 2.5	2.6 ± 2.3
Avg # of utterances per episode	564 ± 54	529 ± 59
Avg # of words/tokens in utterance	7.9 ± 5.4	7.4 ± 5.8
Avg recap duration (s)	104 ± 28	62 ± 20
Avg # of shots in recap	55 ± 12	43 ± 9
Avg # of utterances in recap	33 ± 6	22 ± 5



Acknowledgments: We thank Bank of Baroda for partial travel support, IIIT-H Faculty seed grant and Adobe Research India for student funding.

Qualitative Analysis



Qualitative visualization of TaleSumm predictions on *S06E22* of ***24*** (test set) denoted as *"Ours"*.
GT: Ground-Truth; **F**: A fan site (Fandom) inspired labels; **H**: Human annotated