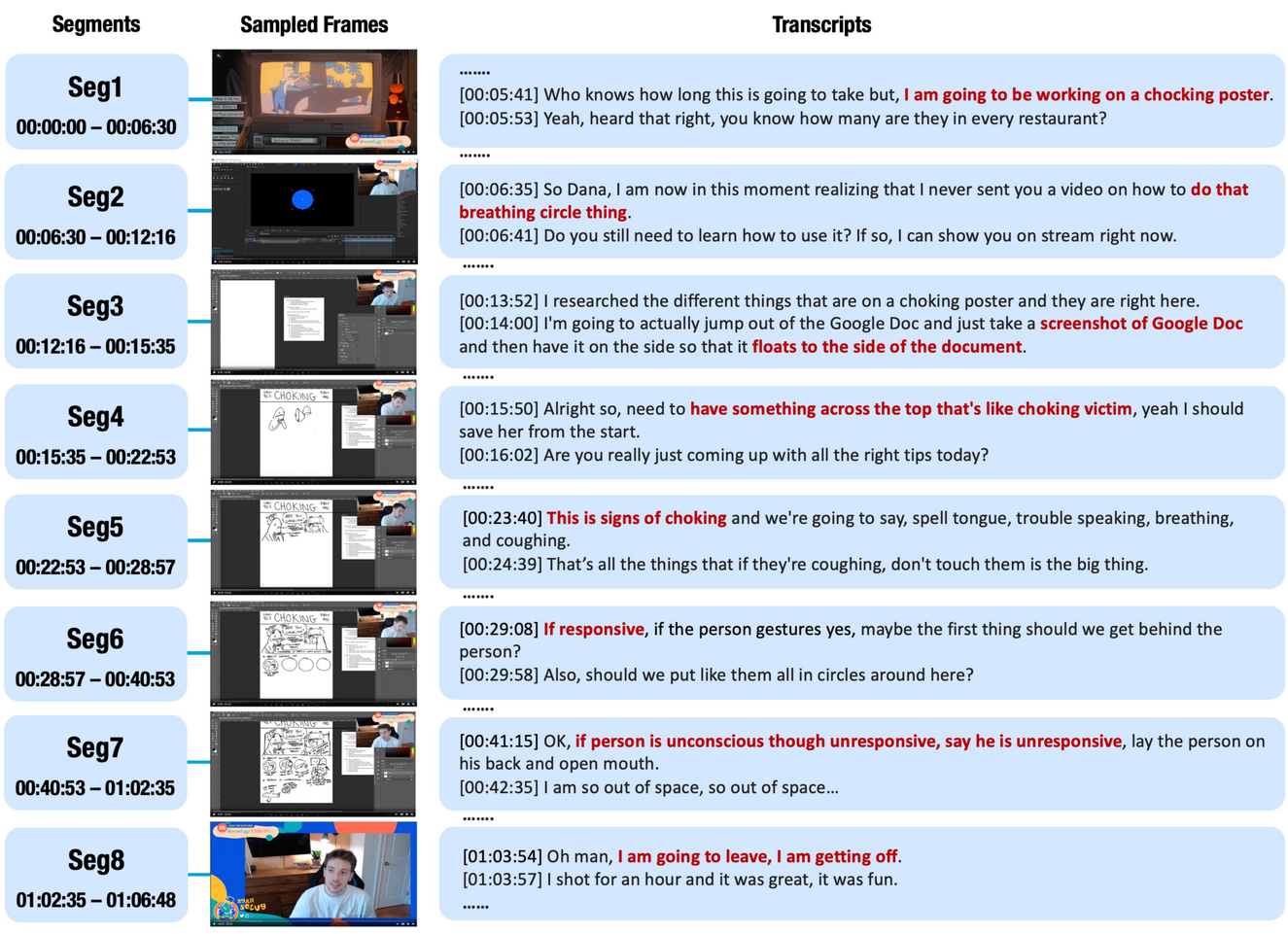
# Multi-modal Video Topic Segmentation with Dual-Contrastive Domain Adaptation

Linzi Xing[1], Quan Tran[2], Fabian Caba[2], Franck Dernoncourt[2], Seunghyun Yoon[2],
Zhaowen Wang[2], Trung Bai[2], Giuseppe Carenini[1]

[1] University of British Columbia [2] Adobe Research

## Motivations




**Tasks similar to video topic segmentation:**

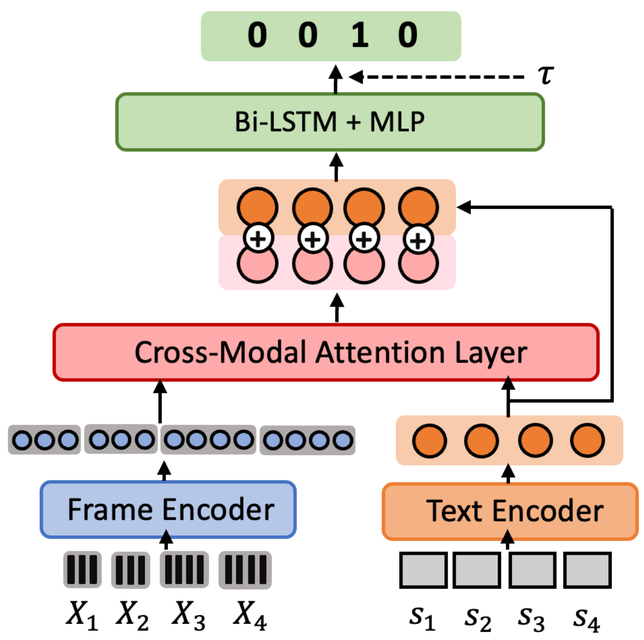
- Shot Segmentation
  - *Uninterrupted, by the same camera.*
- Scene Segmentation
  - *A concept almost exclusively in the movie category.*

**Segments in these tasks are mostly defined by the change of visual features.**

**Challenges for current (shot/scene) approaches:**

- **[Challenge 1]** Livestream videos can have subtle visual changes.
- **[Challenge 2]** A lot of videos are extensively long and from diverse domains.

## Multi-Modal Architecture

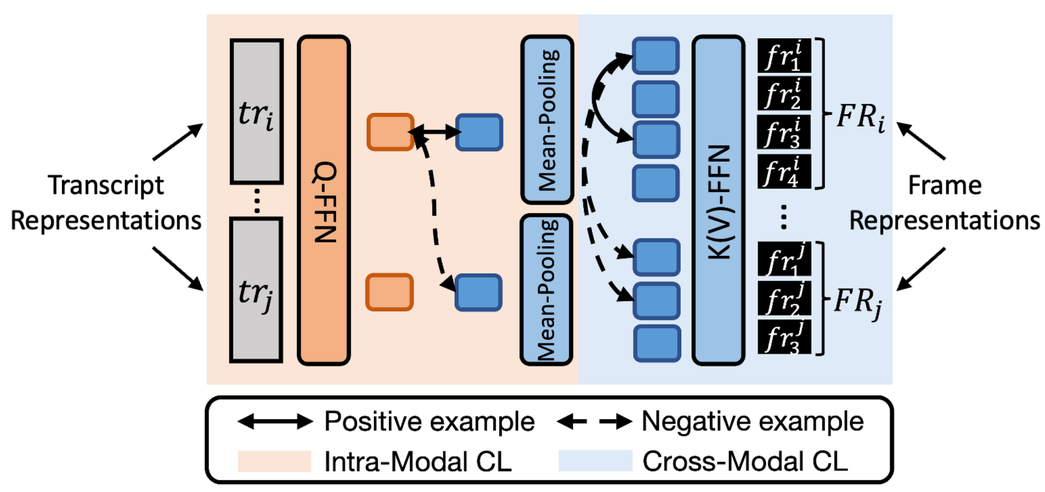



**Multi-Modal Modeling:**

**Transcript (text) + Visual Frames (image)**

- Sequence labeling for text seg. as the base framework.
- Cross-Modal Attention for **text-aware** visual representation.
- Address Challenge 1: Videos with subtle visual changes.

## Dual-Contrastive Adaptation



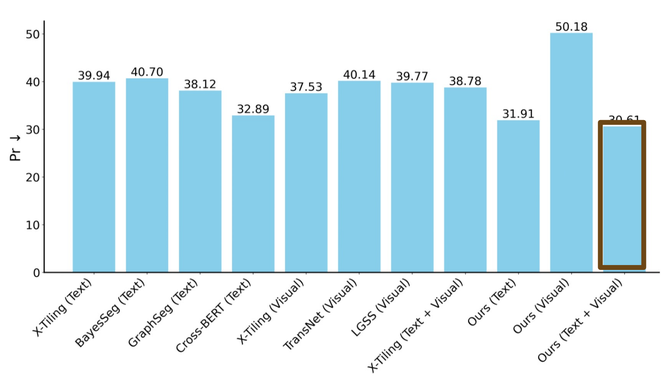

**Dual Contrastive Adaptation:**

- Update model on unlabeled data from target domain.
- Pull the **frames** attached to the same sentence closer and push the ones from different sentences far apart.
- Pull semantically close **sentence-visual pairs** together and push away non-related pairs.
- Address Challenge 2: Long video length from different domains.

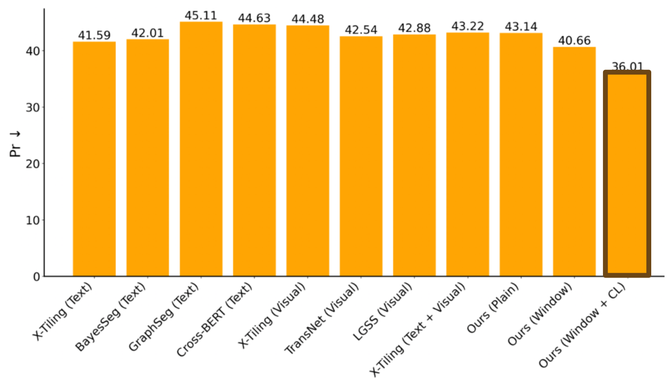
## Experiments

- **Evaluation Metric:** $P_r$ (lower the better)
- **Datasets:**
  - Training: **YouTube** (training set, 5148)
  - Testing: **YouTube** (testing set, 140)
    **BBC Planet Earth** (6)
    **Behance Livestream** (518)
- **[Intra-domain]** Our multi-modal proposal achieves the best performance.
- **[Inter-domain]** The dual-contrastive domain adaptation makes improvements on two target domains.

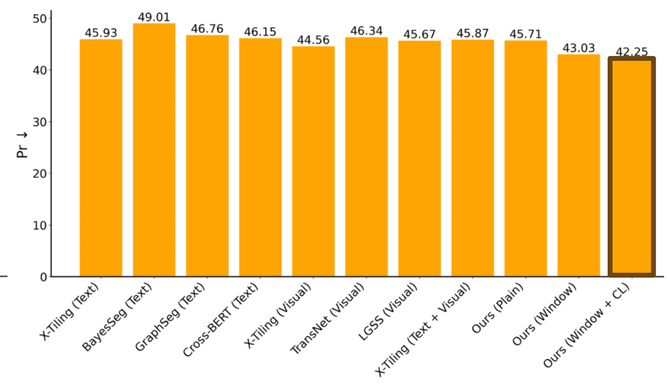
**YouTube**
(intra-domain)




**BBC Planet Earth**
(inter-domain)




**Behance**
(inter-domain)




## Conclusions

- ✓ We can expand supervised segmentation model for monologue to video topic segmentation.
- ✓ Appropriate multi-modal modeling (i.e., feature fusing by cross-modal attention) can improve performance.
- ✓ Unsupervised domain adaptation (dual-contrastive learning) can help robustness on lengthy videos from low-resourced domains.

30th 2024