

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

Linzi Xing¹, Quan Tran², Fabian Caba², Franck Deroncourt², Seunghyun Yoon²,

Zhaowen Wang², Trung Bai², Giuseppe Carenini¹



¹ University of British Columbia

² Adobe Research



Motivations

Segments	Sampled Frames	Transcripts
Seg1 00:00:00 – 00:06:30	 [00:05:41] Who knows how long this is going to take but, I am going to be working on a choking poster. [00:05:53] Yeah, heard that right, you know how many are they in every restaurant?
Seg2 00:06:30 – 00:12:16	 [00:06:35] So Dana, I am now in this moment realizing that I never sent you a video on how to do that breathing circle thing. [00:06:41] Do you still need to learn how to use it? If so, I can show you on stream right now.
Seg3 00:12:16 – 00:15:35	 [00:13:52] I researched the different things that are on a choking poster and they are right here. [00:14:00] I'm going to actually jump out of the Google Doc and just take a screenshot of Google Doc and then have it on the side so that it floats to the side of the document.
Seg4 00:15:35 – 00:22:53	 [00:15:50] Alright so, need to have something across the top that's like choking victim , yeah I should save her from the start. [00:16:02] Are you really just coming up with all the right tips today?
Seg5 00:22:53 – 00:28:57	 [00:23:40] This is signs of choking and we're going to say, spell tongue, trouble speaking, breathing, and coughing. [00:24:39] That's all the things that if they're coughing, don't touch them is the big thing.
Seg6 00:28:57 – 00:40:53	 [00:29:08] If responsive , if the person gestures yes, maybe the first thing should we get behind the person? [00:29:58] Also, should we put like them all in circles around here?
Seg7 00:40:53 – 01:02:35	 [00:41:15] OK, if person is unconscious though unresponsive, say he is unresponsive , lay the person on his back and open mouth. [00:42:35] I am so out of space, so out of space...
Seg8 01:02:35 – 01:06:48	 [01:03:54] Oh man, I am going to leave, I am getting off. [01:03:57] I shot for an hour and it was great, it was fun.

Tasks similar to video topic segmentation:

- Shot Segmentation

Uninterrupted, by the same camera.

- Scene Segmentation

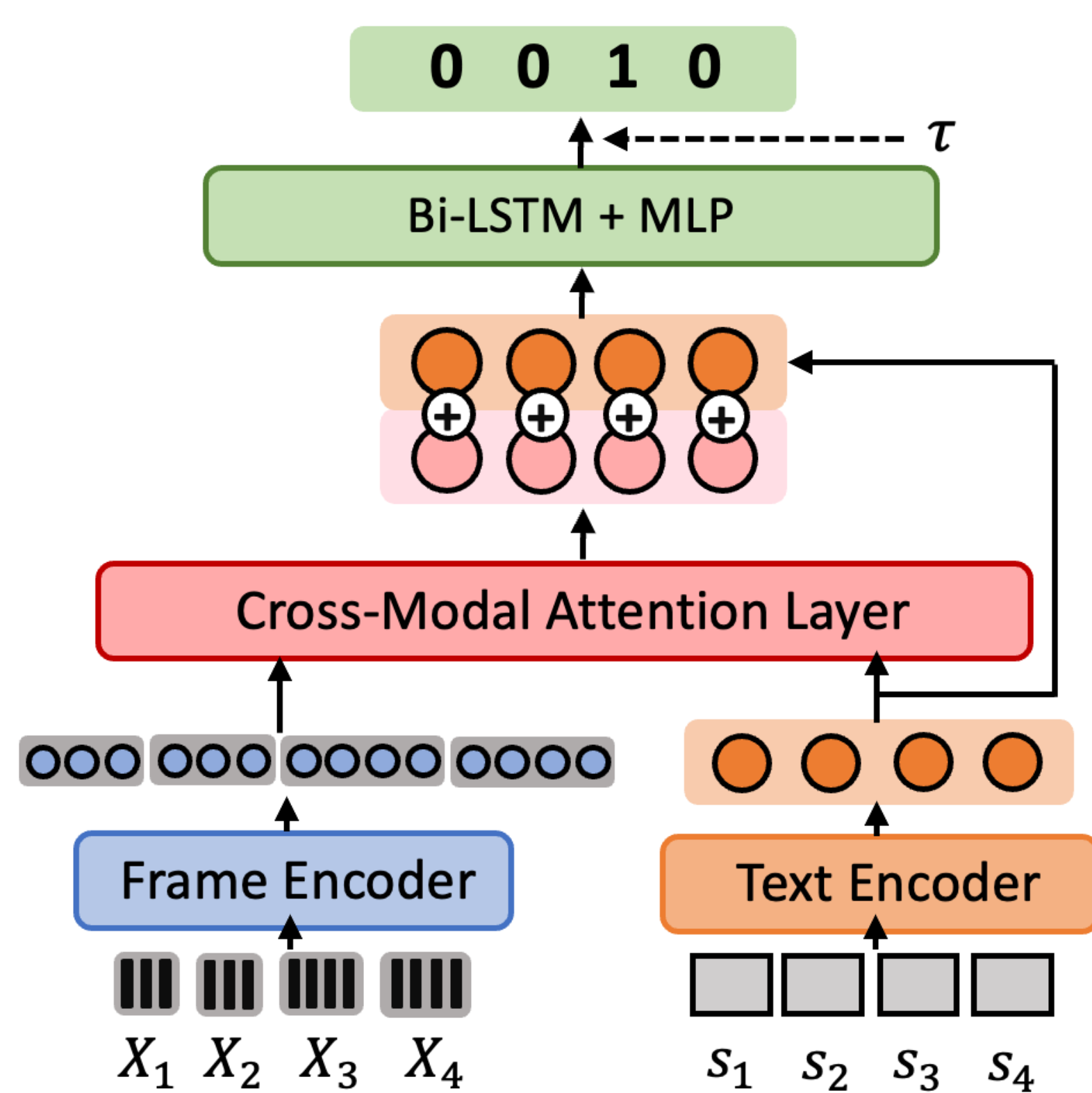
A concept almost exclusively in the movie category.

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

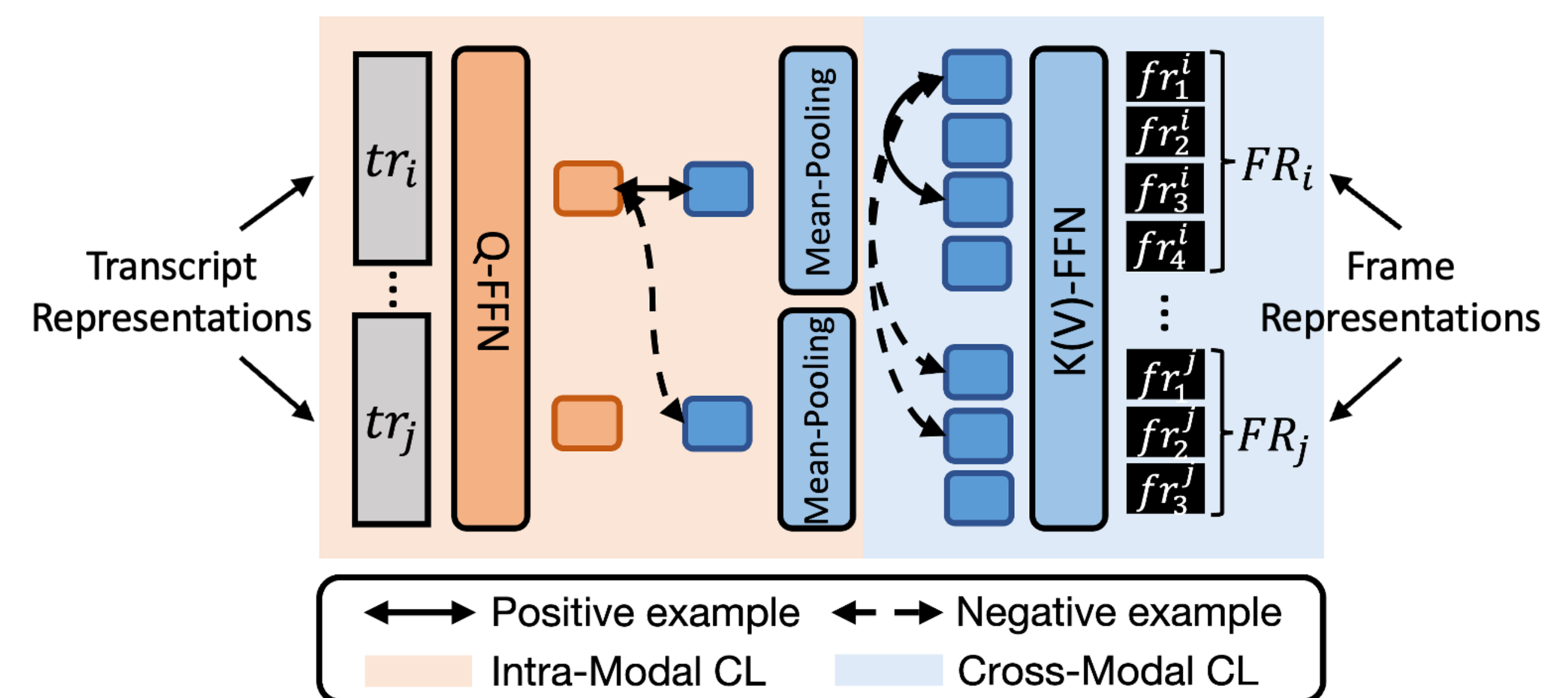


Multi-Modal Modeling:

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.**

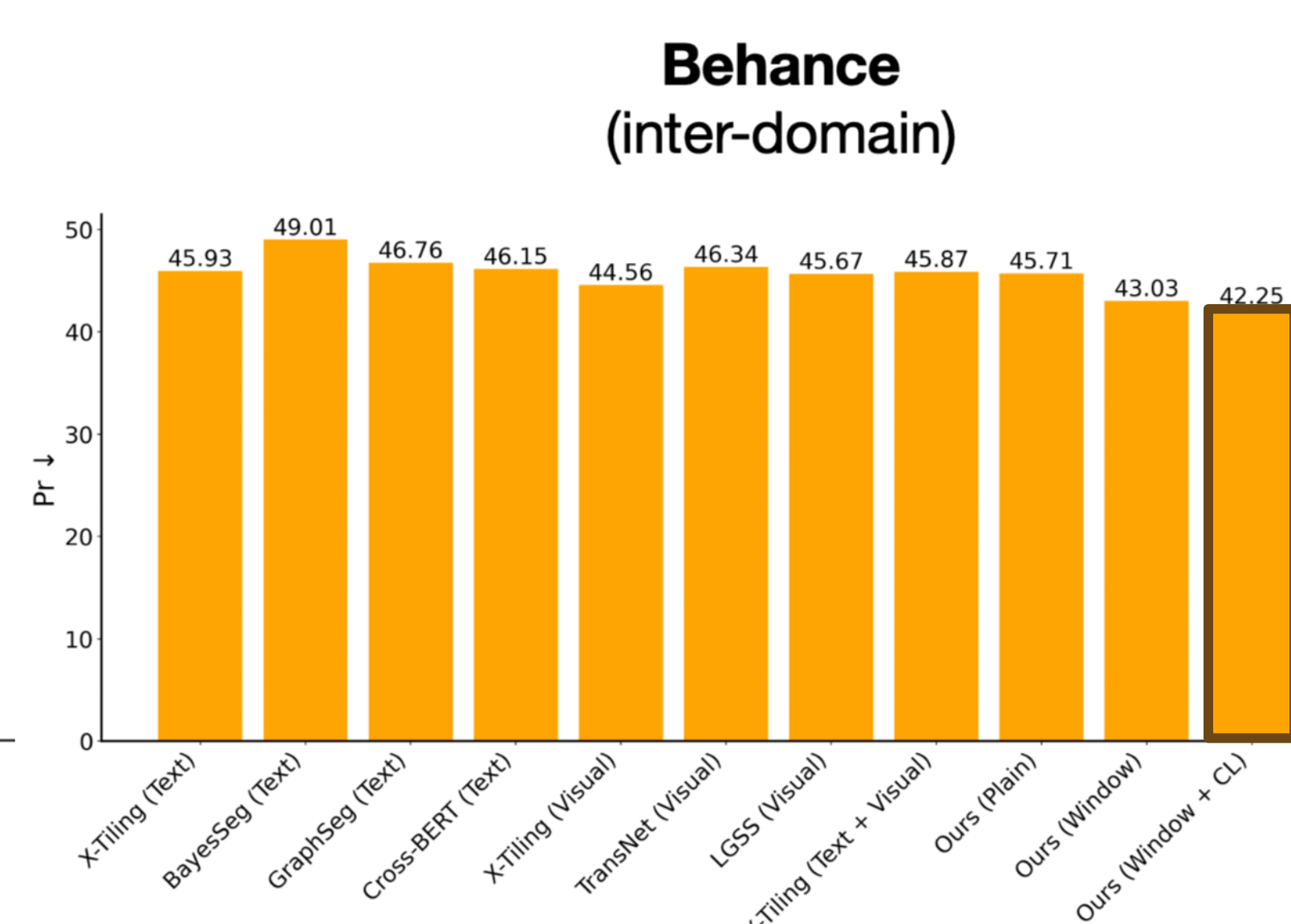
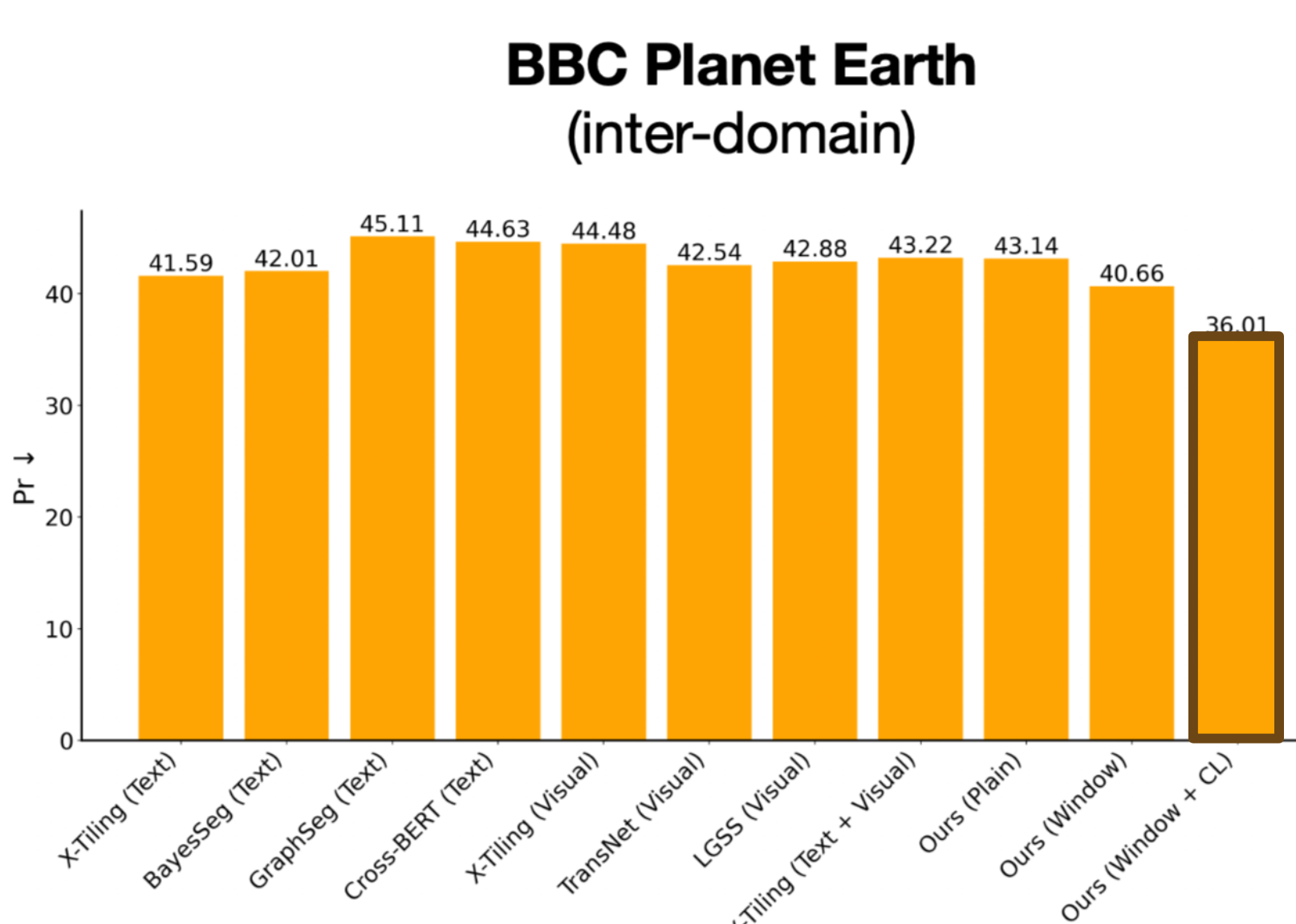
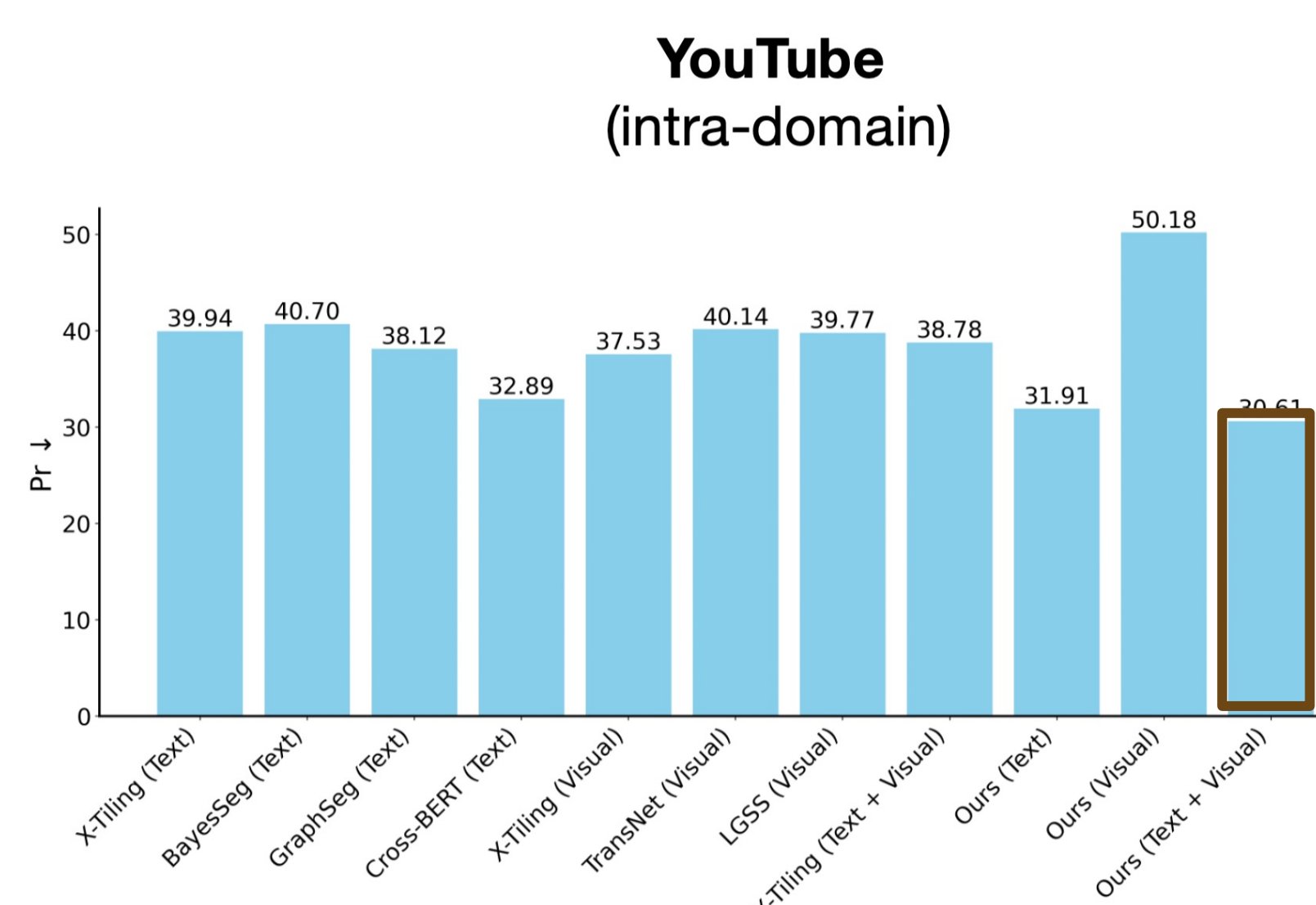
Experiments

- **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.



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.