腾讯优图

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

AI
INIT

NTNU
Norwegian University of
Science and Technology

國立清華大學
NATIONAL TSING HUA UNIVERSITY

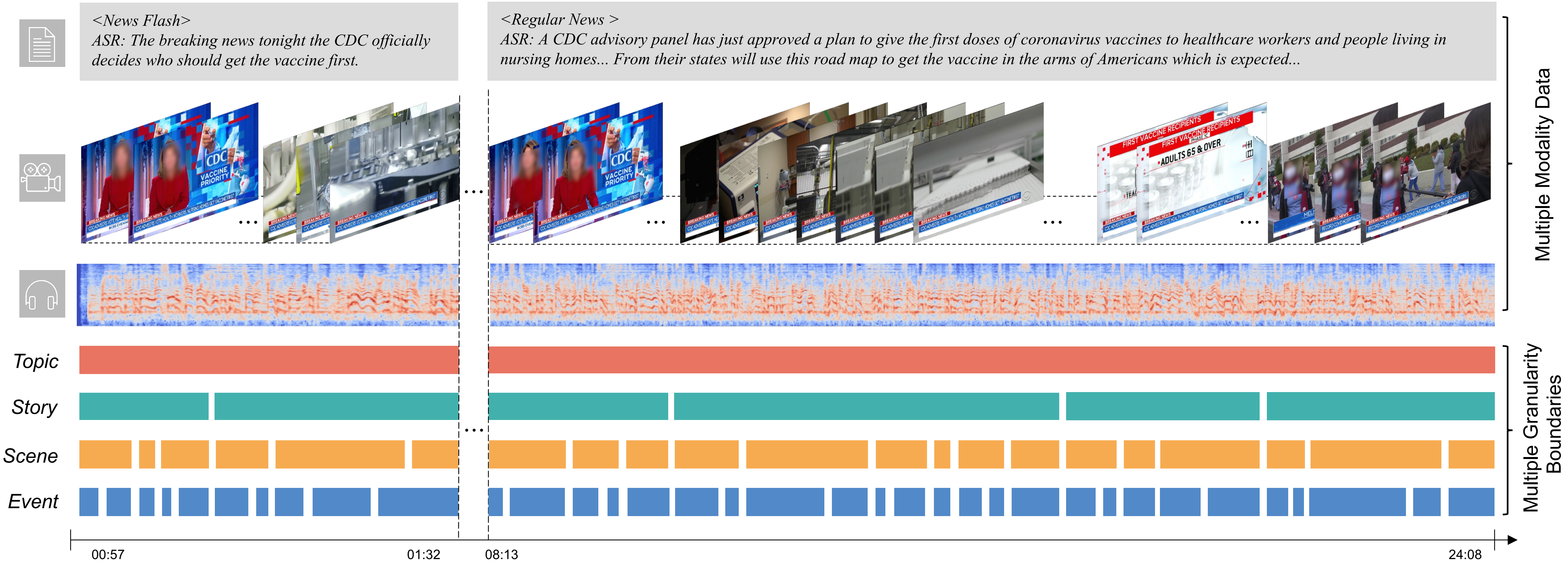
NewsNet: A Novel Dataset for Hierarchical Temporal Segmentation

Haoqian Wu [#1], Keyu Chen [#1], Haozhe Liu [#2], Mingchen Zhuge [#2], Bing Li [*2],Ruizhi Qiao [*1],
Xiujun Shu [1], Bei Gan [1], Liangsheng Xu [1], Bo Ren [1], Mengmeng Xu [2], Wentian Zhang [2],
Raghavendra Ramachandra [3], Chia-Wen Lin [4], Bernard Ghanem [2]

# Joint First Author   * Corresponding Author

[1] Tencent [2] AI Initiative, King Abdullah University of Science and Technology (KAUST)
[3] Norwegian University of Science and Technology (NTNU) [4] National Tsing Hua University (NTHU)

Motivation

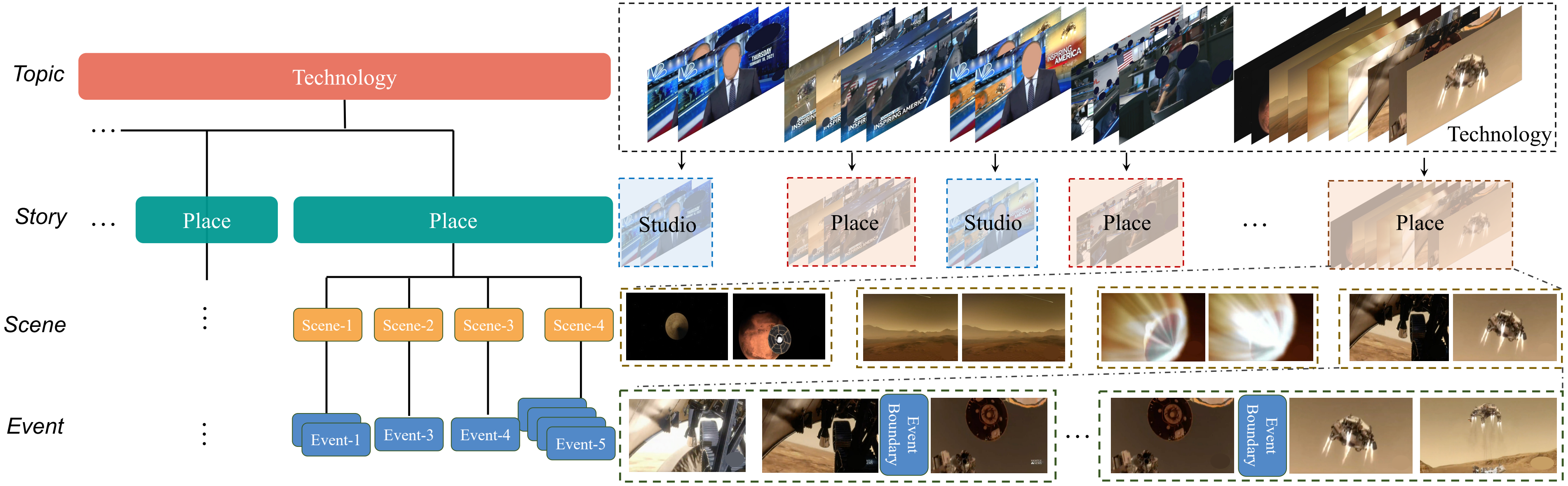


<News Flash>
ASR: The breaking news tonight the CDC officially decides who should get the vaccine first.

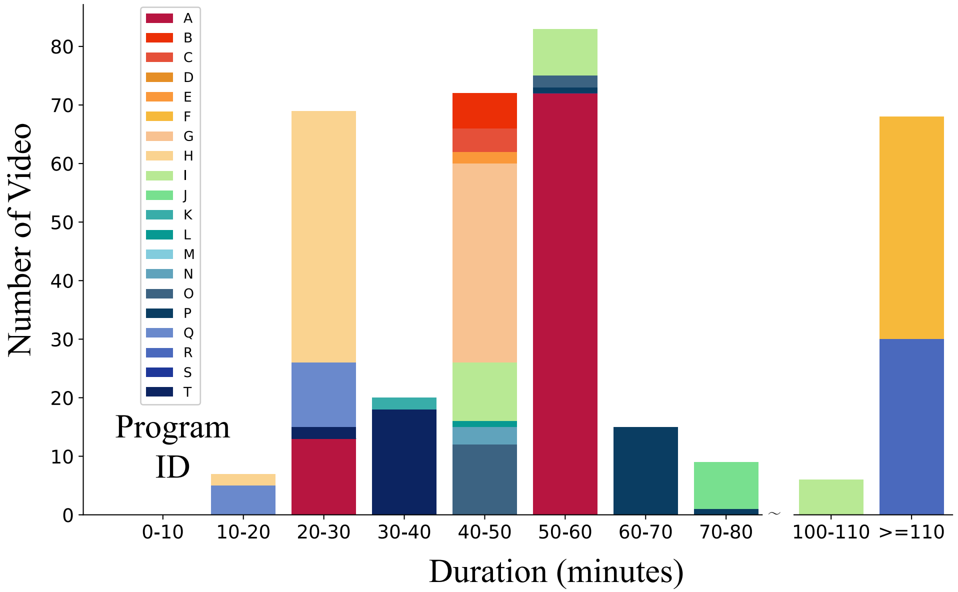
<Regular News >
ASR: A CDC advisory panel has just approved a plan to give the first doses of coronavirus vaccines to healthcare workers and people living in nursing homes... From their states will use this road map to get the vaccine in the arms of Americans which is expected...

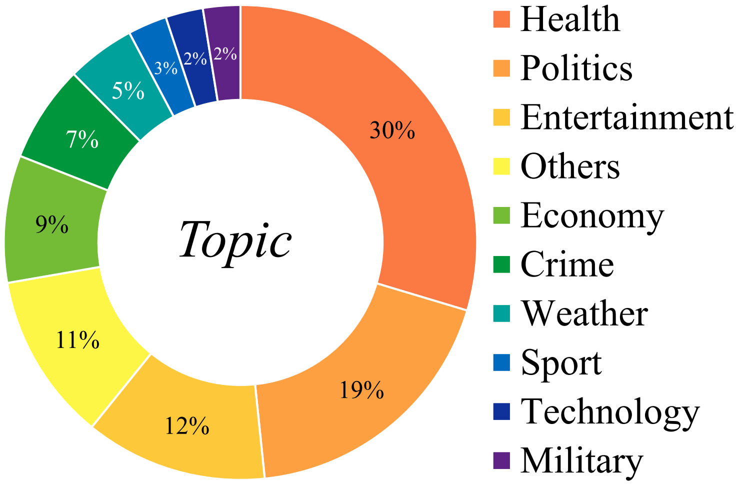
Dataset Summary

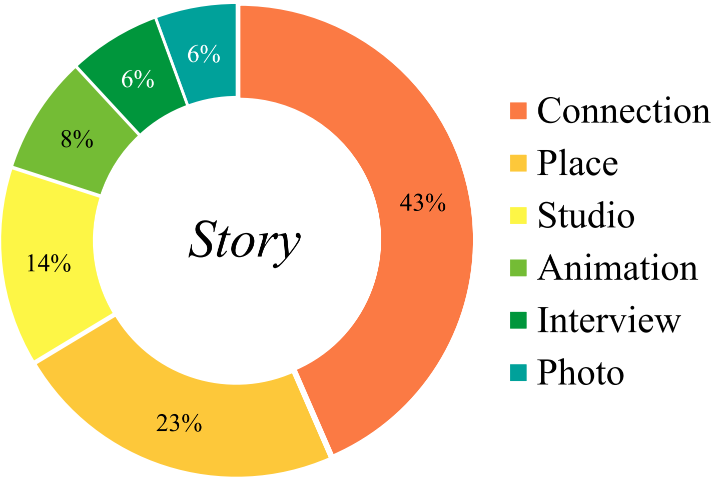
Dataset	# Video	Duration (hours)	Modality	# Annotation(s) per Video				Source
				Topic	Story	Scene	Event	
AVS [75]	197	-	Visual	-	-	-	14.2	Ads
BBC [6]	11	9	Visual	-	-	49.7	-	Doc
OVSD [48]	21	10	Visual	-	-	28.9	-	Generic
Kinetics-GEBD [51]	54,691	152	Audio + Visual	-	-	-	4.9	Action
MovieNet [23] †	1,100	2174	Text + Audio + Visual	-	-	66.0	849.1	Movie
RAI [7]	10	-	Visual	-	-	-	98.7	News
TI-News [35]	477	244	Audio + Visual	-	55.6	-	530.4	News
NewsNet (Ours)	1,000	946	Text + Audio + Visual	8.5	51.6	87.9	654.4	News



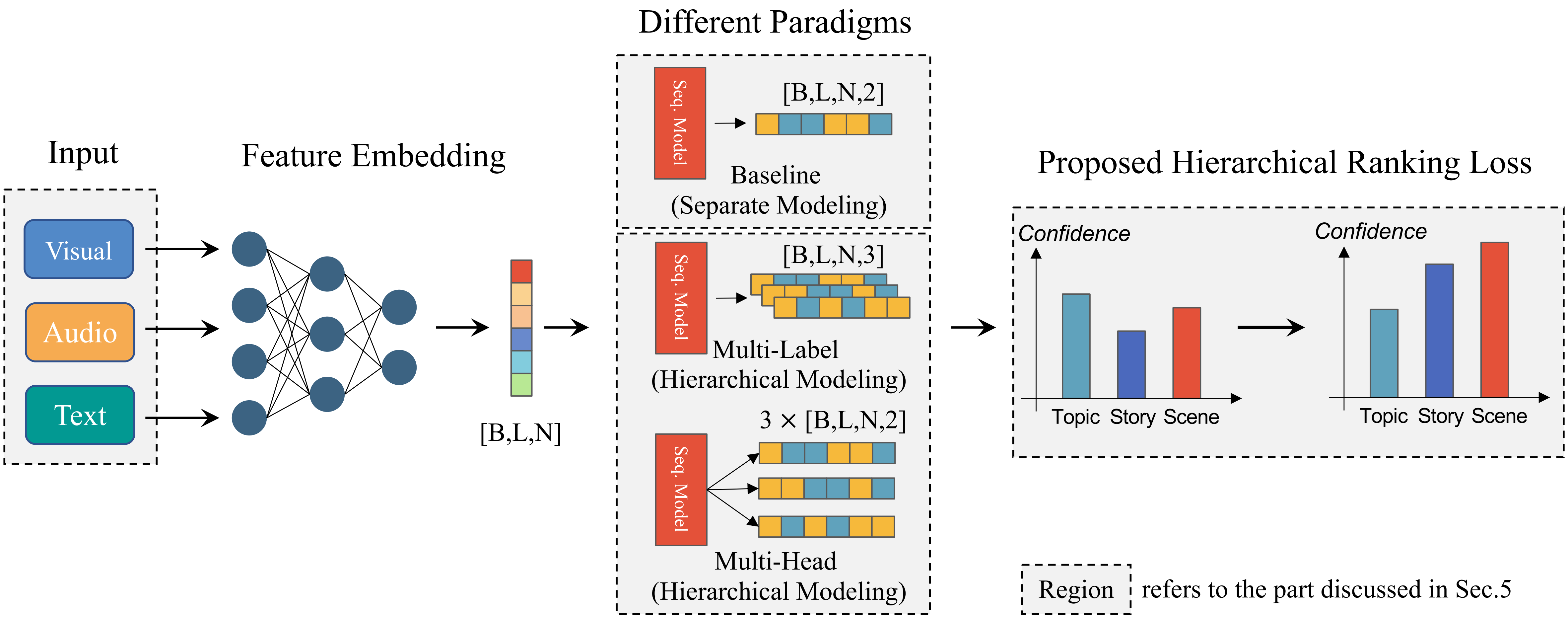
Dataset Summary







Experimental Results



Region refers to the part discussed in Sec.5

Experimental Results

Table 3. In-domain performance by using boundary-free (B.F.) model. The **bolded** values stand for the optimal performances for each task.

Task	Modality	F1 score	Precision	Recall
Scene	V	76.8	76.1	77.5
	A	69.8	66.8	73.2
	T	66.7	56.3	**81.9**
	V+A+T	**78.3**	**80.9**	75.8
Story	V	71.2	72.3	70.0
	A	59.3	57.6	61.1
	T	50.6	57.4	45.2
	V+A+T	**75.4**	**74.7**	**76.2**
Topic	V	62.9	72.4	55.6
	A	58.1	59.4	56.9
	T	39.0	46.5	33.5
	V+A+T	**73.2**	**74.3**	**72.2**

Table 4. Cross-domain setting by using boundary-free (B.F.) model. The **bolded** values stand for the optimal performances for each task.

Task	Modality	Avg. F1 score (std.)	Avg. Precision	Avg. Recall
Scene	V	72.9 (2.1)	70.9	75.2
	A	62.7 (4.0)	59.7	66.6
	T	61.6 (5.0)	52.8	77.0
	V+A+T	**76.0** (2.1)	**74.4**	**77.9**
Story	V	68.5 (2.6)	70.3	66.9
	A	55.7 (3.6)	53.6	59.0
	T	51.1 (3.6)	43.4	65.4
	V+A+T	**72.9** (2.2)	**73.7**	**72.4**
Topic	V	60.6 (4.7)	69.8	53.8
	A	59.0 (5.2)	56.0	62.9
	T	49.8 (5.2)	45.7	55.9
	V+A+T	**72.2** (3.6)	**72.3**	**72.5**

Experimental Results

Table 5. The F1 scores of baselines trained with different levels of annotations on full modalities without our hierarchical ranking loss, where blue and orange indicate the in-domain and cross-domain, respectively. Each row refers to the result corresponding to a single task. Hie. Modeling stands for Hierarchical Modeling while Sep. Modeling is Separate Modeling.

Recipe	Baseline Sep. Modeling	Multi-Label Hie. Modeling	Multi-Head Hie. Modeling
Scene	78.3 / 76.0	79.1 / 76.5	**79.9** / **76.9**
+ Story	75.4 / 72.9	**75.4** / **74.7**	74.2 / 74.0
Scene	78.3 / 76.0	**79.8** / 76.4	79.5 / **76.5**
+ Topic	**73.2** / 72.2	70.5 / 72.8	70.9 / **73.0**
Story	75.4 / 72.9	**76.2** / **74.3**	75.4 / 73.9
+ Topic	73.2 / 72.2	**77.3** / 73.2	75.2 / **73.5**
Scene	78.3 / 76.0	77.4 / **76.8**	**79.8** / **76.8**
+ Story	**75.4** / 72.9	74.3 / **74.3**	74.5 / 73.7
+ Topic	73.2 / 72.2	74.3 / **72.6**	**76.6** / 70.4

Table 6. The F1 scores of the methods with or without hierarchical ranking loss under the in-domain / cross-domain setting on full modalities. Hie. stands for Hierarchical Modeling and Sep. refers to Separate Modeling.

Method	Scene	Story	Topic
Baseline (Sep.)	78.3 / 76.0	75.4 / 72.9	73.2 / 72.2
Multi-Label (Hie.)	77.4 / 76.8	74.3 / **74.3**	74.3 / 72.6
Multi-Label w/ Hie. Loss (Hie.)	**79.6** / **76.9**	**74.4** / 73.5	**77.8** / **73.1**
Multi-Head (Hie.)	79.8 / 76.8	74.5 / 73.7	**76.6** / 70.4
Multi-Head w/ Hie. Loss (Hie.)	**80.3** / **76.9**	**76.3** / **74.6**	76.5 / **73.2**

Conclusion

1. We propose a novel large-scale dataset NewsNet for long-form video structure understanding. This dataset is derived from 900+ hours of video and annotated with 4 hierarchical levels of semantics.

2. NewsNet provides dense annotations and multi-modal information, promoting diverse benchmarks: separate/hierarchical temporal video segmentation in scene/story/topic levels, as well as other common tasks like classification, video localization/grounding, and highlight detection.

3. We formulate a new benchmark, i.e., hierarchical modeling in the temporal segmentation task, which needs a single model to predict segments of multiple hierarchical levels. Based on the empirical study, we bring insights into how hierarchical modeling potentially benefits the temporal video segmentation task, which was almost never discussed.

Thank you for listening!