Align and Attend: Multimodal Summarization with Dual Contrastive Losses

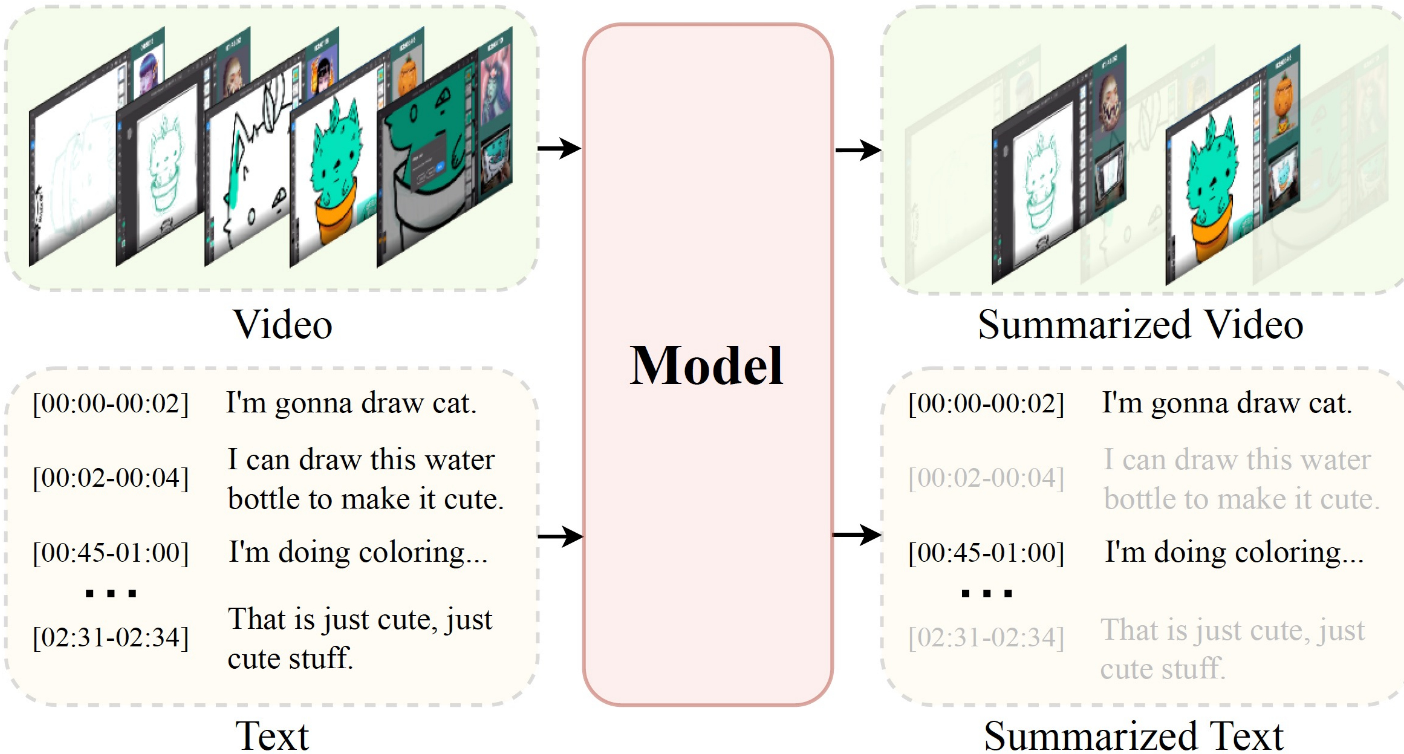
Bo He[1], Jun Wang[1], Jielin Qiu[2], Trung Bui[3], Abhinav Shrivastava[1], Zhaowen Wang[2]

[1] University of Maryland, College Park [2] Carnegie Mellon University [3] Adobe Research

Poster: WED-PM-240

Website and Code: https://boheumd.github.io/A2Summ

Task Definition

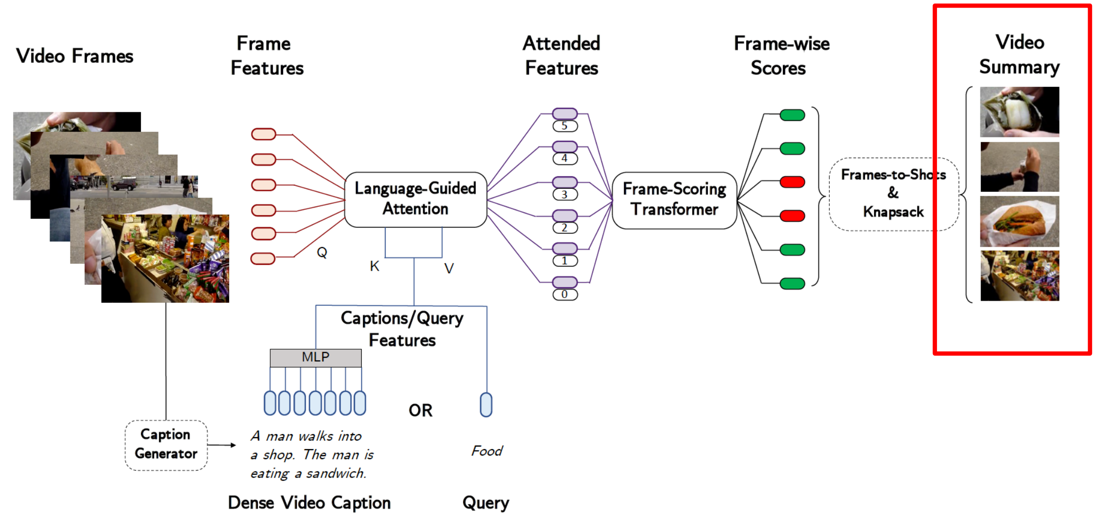


Multimodal Summarization with Multimodal Output

Prior Work

✕ Multimodal Output

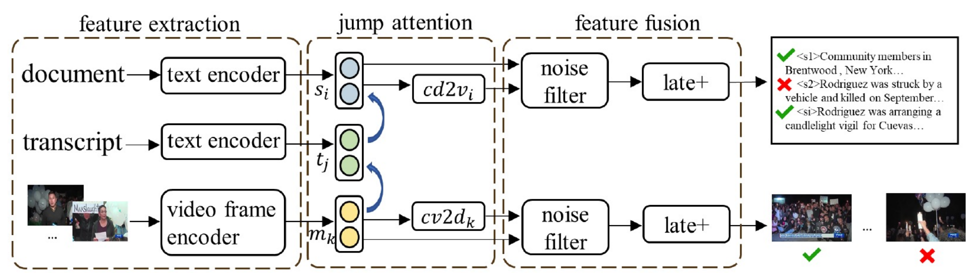
- The additional modality acts as auxiliary input information.



Clip-It! *[Narasimhan et al., NeurIPS'21]*

✕ Alignment across Multimodal

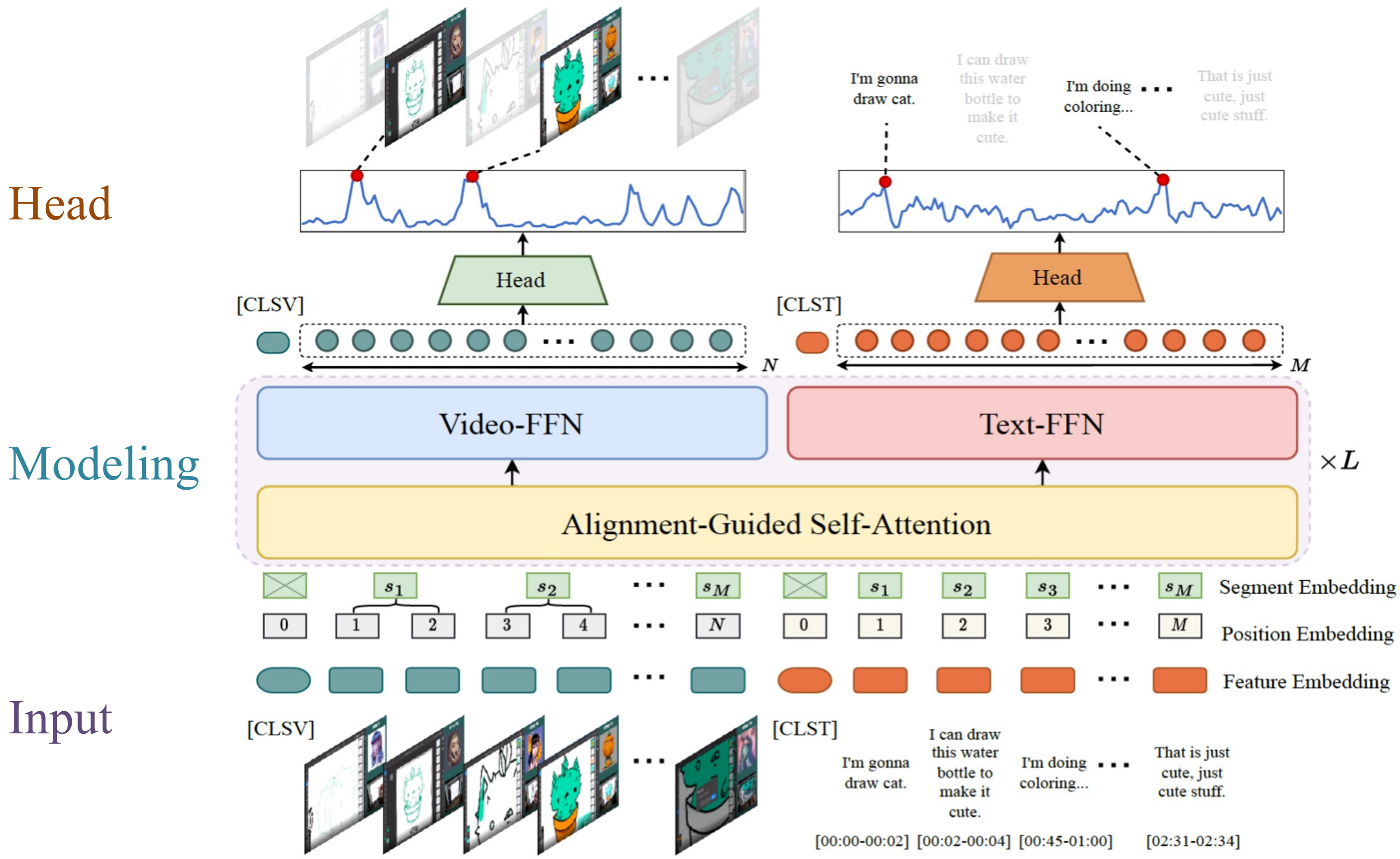
- Time correspondence between different modalities is ignored.



M2SM *[Fu et al., NAACL'21]*

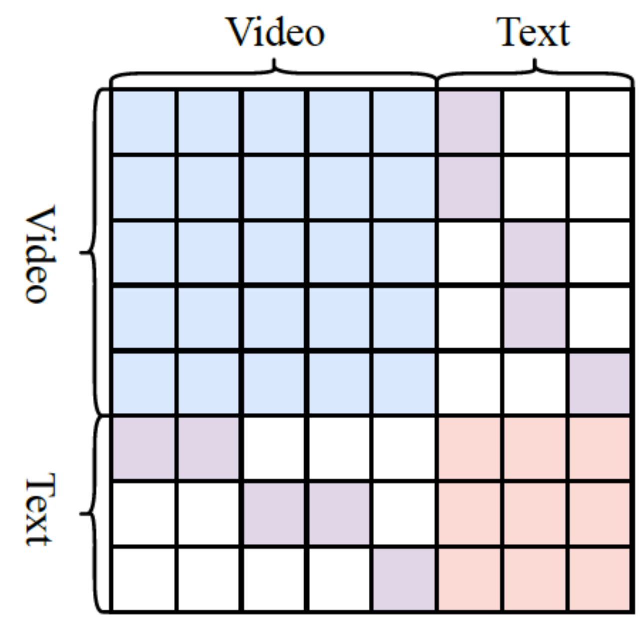
Motivation: Align multimodal input and exploit intrinsic cross-modality relationship.

Model Architecture



Alignment-guided Self-attention

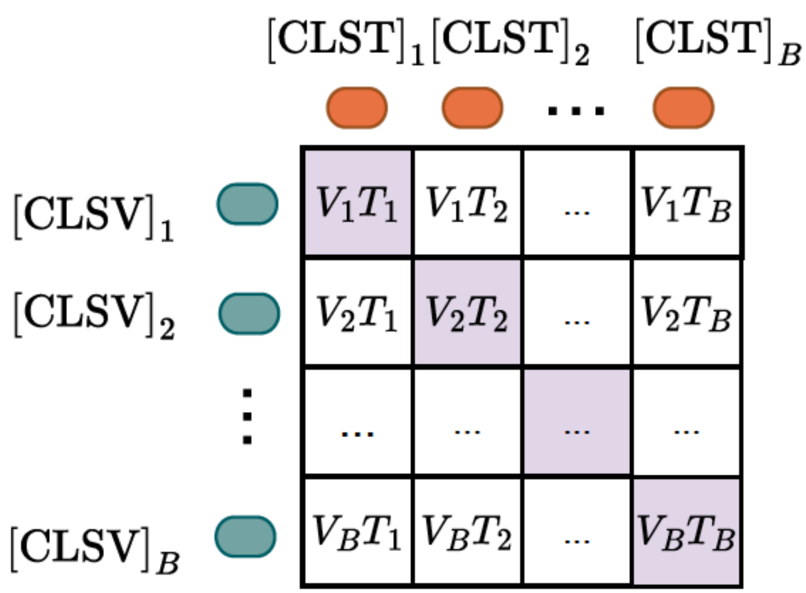
- **Goal:** Exploit *time correspondences* between video and text sequences.
- **Motivation:** Untrimmed videos and text sentences contain irrelevant backgrounds.
- **Solution:** Video frames and text sentences from the *same time window* can attend to each other.



Attention Mask in Self-attention Layer

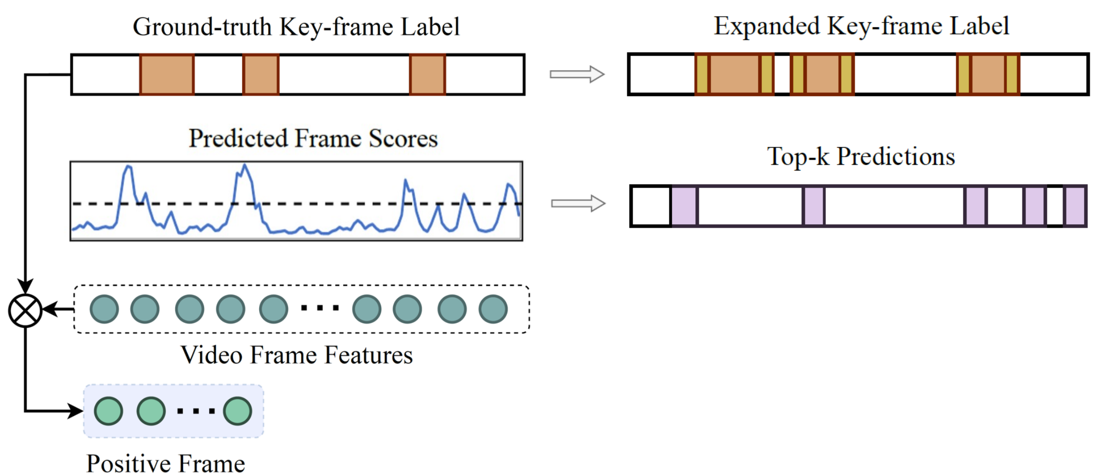
Inter-Sample Contrastive Loss

- **Goal:** Utilize *intrinsic* relationships between input multimodal data pair.
- **Motivation:** Paired video and text sample shares mutual information.
- **Solution:** Maximize the cosine similarity of the video and text embedding from B real pairs in the batch while minimizing the similarity from $B^2 - B$ incorrect pairs.



Intra-Sample Contrastive Loss

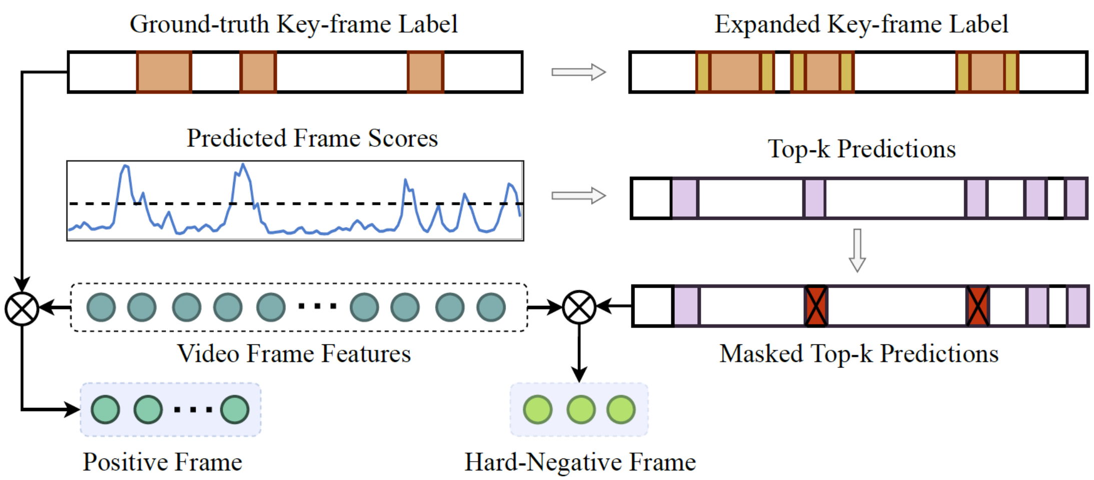
- **Goal:** Model *fine-grained* cross-modality information.
- **Motivation:** Human-annotated key-frames and key-sentences reveal the same semantic meanings. (e.g., cooking recipe video)



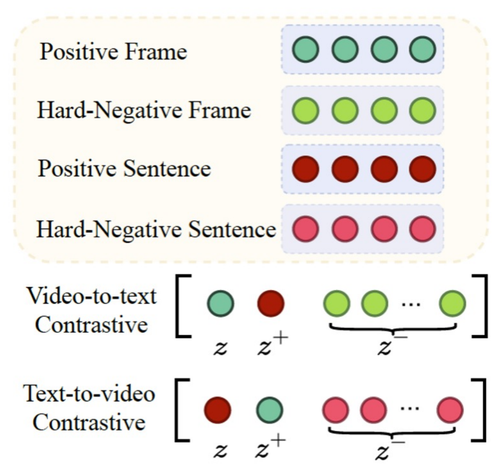
Contrastive Pair Selection

Intra-Sample Contrastive Loss

- **Goal:** Model *fine-grained* cross-modality information.
- **Motivation:** Human-annotated key-frames and key-sentences reveal the same semantic meanings. (e.g., cooking recipe video)



Contrastive Pair Selection



Contrastive Loss

Ablation Studies

Inputs	Align.	Inter.	Intra.	F1	τ	ρ
Video-Only				49.8	0.070	0.084
Multimodal				50.5	0.083	0.096
	✓			51.5	0.089	0.104
	✓	✓		52.5	0.905	0.110
	✓		✓	54.0	0.102	0.121
	✓	✓	✓	**55.0**	**0.108**	**0.129**

+5.2 F1 score

Contribution of each component

- Align: Alignment-guided self-attention
- Inter: Inter-sample Contrastive Loss
- Intra: Intra-sample Contrastive Loss

Comparison with SOTA

Category	Method	CNN			Daily Mail			
		R-1	R-2	R-L	R-1	R-2	R-L	Cos(%)
Video	VSUMM [30]	–	–	–	–	–	–	68.74
	DR-DSN [11]	–	–	–	–	–	–	68.69
	CLIP-It [2]	–	–	–	–	–	–	69.25
Text	Lead3 [18]	–	–	–	41.07	17.87	30.90	–
	SummaRuNNer [20]	–	–	–	41.12	17.92	30.94	–
	NN-SE [19]	–	–	–	41.22	18.15	31.22	–
Multimodal	MM-ATG [1]	26.83	8.11	18.34	35.38	14.79	25.41	69.17
	Img+Trans [5]	27.04	8.29	18.54	39.28	16.64	28.53	–
	TFN [82]	27.68	8.69	18.71	39.37	16.38	28.09	–
	HNNattTI [4]	27.61	8.74	18.64	39.58	16.71	29.04	68.76
	M^{2}SM [6]	27.81	8.87	18.73	41.73	18.59	31.68	69.22
Ours	Video-only	–	–	–	–	–	–	69.30
	Text-only	29.39	10.85	26.11	42.77	19.19	34.60	–
	A2Summ	**30.82**	**11.40**	**27.40**	**44.11**	**20.31**	**35.92**	**70.20**

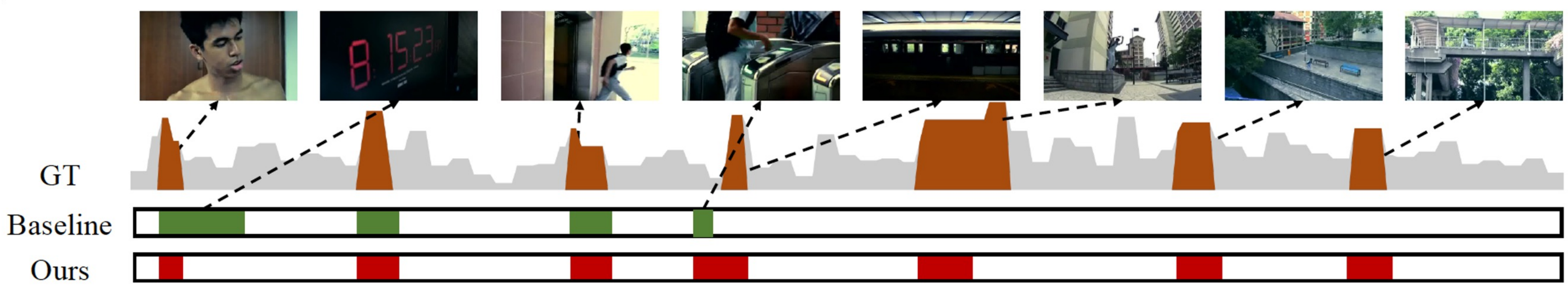
CNN and Daily Mail Datasets

Comparison with SOTA

Method	SumMe			TVSum		
	F1	τ	ρ	F1	τ	ρ
Random [76]	41.0	0.000	0.000	57.0	0.000	0.000
Human [76]	54.0	0.205	0.213	54.0	0.177	0.204
DR-DSN [11]	42.1	–	–	58.1	0.020	0.026
HSA-RNN [45]	42.5	0.064	0.066	44.1	0.082	0.088
CSNet [33]	48.6	–	–	58.5	0.025	0.034
VASNet [84]	49.7	–	–	61.4	–	–
DSNet-AB [14]	50.2	0.051	0.059	62.1	0.108	0.129
DSNet-AF [14]	51.2	0.037	0.046	61.9	0.113	0.138
RSGN [16]	45.0	0.083	0.085	60.1	0.083	0.090
CLIP-It [2]	51.6	–	–	**64.2**	0.108	0.147
iPTNet [17]	<u>54.5</u>	<u>0.101</u>	<u>0.119</u>	63.4	<u>0.134</u>	<u>0.163</u>
A2Summ	**55.0**	**0.108**	**0.129**	<u>63.4</u>	**0.137**	**0.165**

SumMe and TVSum Datasets

Visualization



Description: *"A boy woke up late, took the metro and then ran to school."*

A2Summ generates summaries which cover important segments with more accurate temporal boundaries.

BLiSS Dataset

- A **large-scale** multimodal summarization dataset focused on the livestream videos and transcripts.

	SumMe	TVSum	CNN	Daily Mail	StreamHover	BLiSS
Number of Data	25	50	203	1970	5421	13303
Total Video Duration (Hours)	1.0	3.5	7.1	44.2	452	1109
Total Number of Text Tokens	–	–	0.2M	1.3M	3.1M	5.5M
Avg. Video Summary Length	44	70	–	2.9	–	10.1
Avg. Text Summary Length	–	–	29.7	59.6	79	49

Statistics Comparison

Thank You!

For more details, please visit

Poster# 240 at

21-Jun-23, 4:30pm-6:30pm

Website and Code:

https://boheumd.github.io/A2Summ