https://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *INFORMS Journal on Computing*

ISSN 0899-1499, EISSN 1526-5528

Online Supplement for "CAAC: Co-attentive Actionability Classification for Assessing Patient Education Videos"

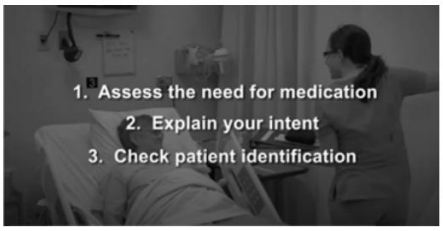
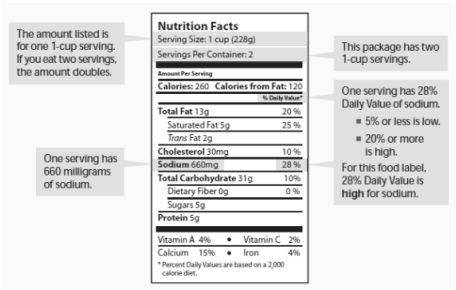
Krishna Pothugunta,[a] Xiao Liu,[b] Anjana Susarla,[c] Rema Padman,[d]

[a]Department of Information Technology, Analytics and Operations, Mendoza College of Business, University of Notre Dame, kpothugu@nd.edu; [b]Department of Information Systems, W. P. Carey School of Business, Arizona State University, xiao.liu.10@asu.edu; [c]Department of Accounting and Information Systems, Michigan State University, asusarla@msu.edu; [d]Heinz College of Information Systems and Public Policy, Carnegie Mellon University, rpadman@andrew.cmu.edu

Abstract. This document contains the online supplemental materials for the submission "CAAC: Co-attentive Actionability Classification for Assessing Patient Education Videos."

Online Supplement A: Actionability Annotation

Table A.1 A Sample Video on Actionability Annotation

1. The material clearly identifies at least one action the user can take.	2. The material addresses the user directly when describing actions.
	
3. The material breaks down any action into manageable, explicit steps.	**4. The material explains how to use charts, graphs, tables, or diagrams to take actions.**
	

URL: https://www.youtube.com/watch?v=prFYht1NwmU

Online Supplement B: Notations

Table B.1 Summary of notations.

Notation	Description
a	Index of a video in the dataset
y_a	Actionability label for video a (1 = actionable, 0 = not actionable)
d	Dimensionality of hidden embeddings (text and vision)
$\mathcal{S}_a$	Set of all visually coherent segments in video a
R_a	Number of segments in video a
$\mathcal{S}_{a,r}$	r-th segment in video a
$[\tau_{a,r}^{\text{start}}, \tau_{a,r}^{\text{end}}]$	Start and end timestamps of segment r
F_a	Set of all sampled frames across all segments
$J_{a,r}$	Number of sampled frames in segment r
$f_{a,r,j}$	j-th frame sampled (at 1 fps) from segment r of video a
$t_{a,r,j}$	Timestamp associated with frame $f_{a,r,j}$
$\mathbf{P}_a$	Set of all frame patches across the video
$N_{a,r,j}$	Number of visual patches extracted from frame $f_{a,r,j}$
$\mathbf{p}_{a,r,j,n}$	n-th patch of frame $f_{a,r,j}$
$(w_{a,i}, t_{a,i}^{w})$	Word $w_{a,i}$ and its ASR timestamp $t_{a,i}^{w}$
$W_{a,r}$	Words whose timestamps fall in segment interval $[\tau_{a,r}^{\text{start}}, \tau_{a,r}^{\text{end}}]$
I_a	Total number of words in the transcript of video a
Δt	Frame sampling interval (1 second)
$W_{a,r,j}$	Words spoken during the 1-second window of frame $f_{a,r,j}$
$W_{a,r,j} \subseteq W_{a,r}$	Frame-level word set is a subset of the segment-level word set
$I_{a,r,j}$	Number of words in the frame-level word set $W_{a,r,j}$
$\mathbf{X}_{a,r,j}$	Textual encoder token embeddings for $W_{a,r,j}$ $((I_{a,r,j}+2)\times d)$
$\mathbf{P}_{a,r,j}$	ViT patch embeddings for frame $f_{a,r,j}$ $((N_{a,r,j}+1)\times d)$
U_p, U_t	Linear projection matrices for visual and textual features
$l_{a,r,j,i,n}^{T\to F}$	Text-to-frame co-attention logit for token i, patch n
$l_{a,r,j,n,i}^{F\to T}$	Frame-to-text co-attention logit for patch n, token i
$A_{a,r,j,i,n}^{T\to F}$	Text-to-frame attention weight for token i, patch n
$A_{a,r,j,n,i}^{F\to T}$	Frame-to-text attention weight for patch n, token i
$\mathbf{C}_{a,r,j}^{T\to F}$	Frame-conditioned text co-attention embedding
$\mathbf{C}_{a,r,j}^{F\to T}$	Word-conditioned frame co-attention embedding
$\mathbf{A}_{a,r,j}^{\text{proj}}$	Projection matrix aligning text sequence length to patch sequence length
$\hat{\mathbf{C}}_{a,r,j}^{T\to F}$	Projected text-to-frame co-attention embedding
$v_{\text{fw}}, b_{\text{fw}}$	Parameters of the frame–words fusion gate
$\mathbf{CA}_{a,r,j}$	Fused frame–words co-attention embedding for frame $f_{a,r,j}$
$\mathbf{CA}_{a,r}$	Matrix of frame-level embeddings in segment r $(J_{a,r}\times d)$
$v_{\text{seg}}, b_{\text{seg}}$	Parameters of segment-level temporal attention
$\boldsymbol{\lambda}_{a,r}$	Attention weights over frames in segment r
$\widetilde{\mathbf{CA}}_{a,r}$	Segment-level embedding for segment $\mathcal{S}_{a,r}$

Table B.1 Summary of notations (contd.).

Notation	Description
$\phi_{a,r}^{T\to F}(j)$	Text-to-frame temporal attention distribution over frames in segment r
$\phi_{a,r}^{F\to T}(j)$	Frame-to-text temporal attention distribution over frames in segment r
$H_{a,r}^{T\to F}$	Entropy of $\phi_{a,r}^{T\to F}$ within segment r
$H_{a,r}^{F\to T}$	Entropy of $\phi_{a,r}^{F\to T}$ within segment r
$\mathbf{CA}_a$	Matrix of segment-level embeddings in video a ($R_a \times d$)
$v_{\text{vid}}, b_{\text{vid}}$	Parameters of video-level segment attention
$\boldsymbol{\gamma}_a$	Attention weights over segments in video a
$\widetilde{\mathbf{CA}}_a$	Video-level embedding obtained by aggregating segment embeddings
$\bar{H}_a^{T\to F}$	Video-level mean entropy of text-to-frame temporal distributions
$\bar{H}_a^{F\to T}$	Video-level mean entropy of frame-to-text temporal distributions
$\mathcal{L}_{\text{cls}}(\widetilde{\mathbf{CA}}_a, y_a)$	Classification loss (binary cross-entropy) for video a
$\mathcal{L}$	Total training loss with entropy regularization
α, β	Hyperparameters controlling strength of entropy penalties

Online Supplement C: CAAC Framework (Segment and Video-Level Aggregation)

Figure C.1 Segment-Level (across frames) Aggregation: Level 2.

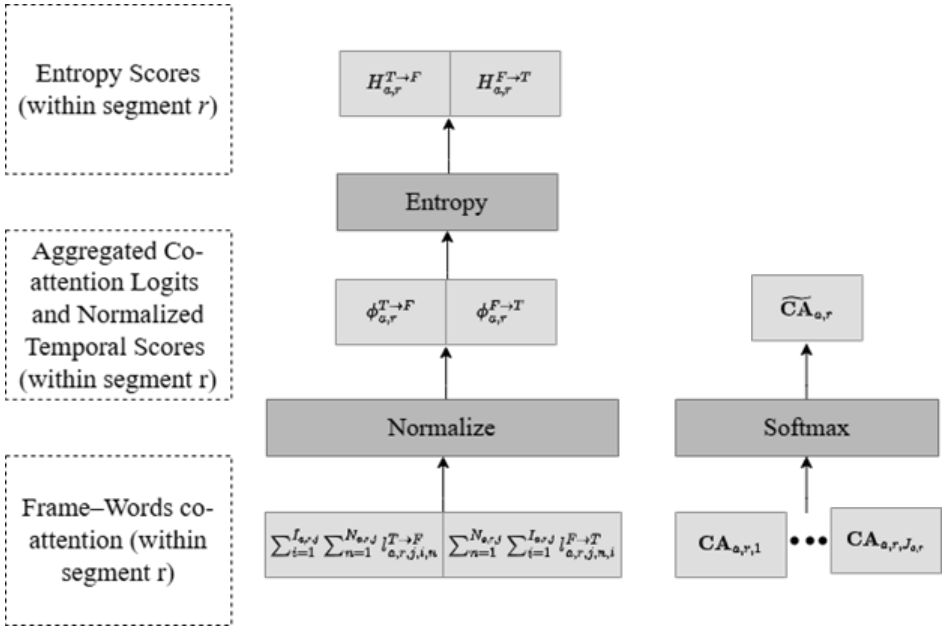
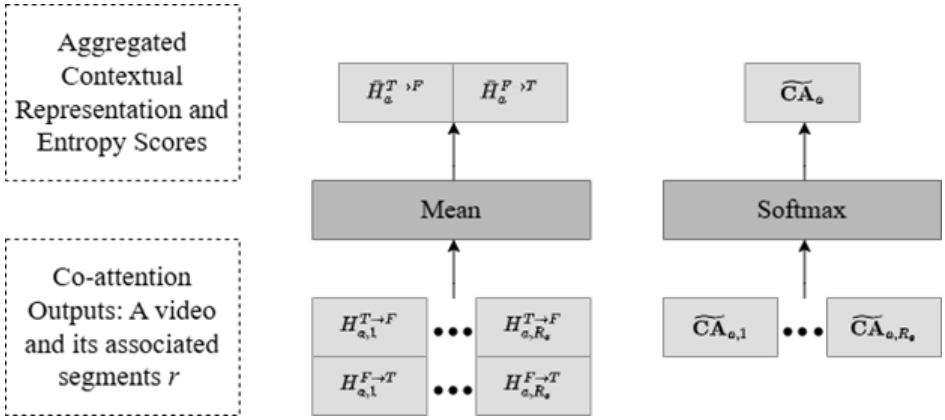


Figure C.2 Video-Level (across segments) Aggregation: Level 3.



Online Supplement D: Transcript Retrieval and Summary Statistics

Figure D.1 Retrieval of Video Transcripts.

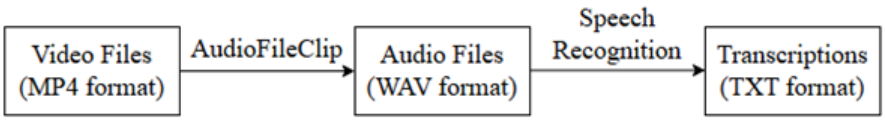


Table D.1 Sample Video Transcript Breakdown

Sentence	Start (s)	End (s)
Sentence-Based		
Diabetes connections is brought to you	1	4
by Animus corporation providing insulin	4	9
delivery products for people living with	9	12
diabetes and part of the OneTouch family	12	17
Screen-Change		
Diabetes connections is brought to you by Animus corporation providing insulin	1	9
delivery products for people living with diabetes and part of the OneTouch family	7	17

Table D.2 Summary Statistics

Variable	Mean	S.D.	Min.	Max.
Sentences per Video	50	39	1	467
Frames per Video	290	194	28	2338
Frames per Segment	7	3	2	50

Online Supplement E: Inputs, Attention Variation and Alternate Encoders

Table E.1 Inputs and Attention Variation (Sentence-based Alignment)

Model	Hyper-parameters	AUC	F1-Score	Precision	Recall
Base Inputs					
Frame-only	*lr = 0.01* *wd = 0.0001*	0.708***	0.571***	0.707***	0.643***
Text-only	*lr = 0.01* *wd = 0.001*	0.764***	0.689***	0.721***	0.667***
Attention Variants					
CAAC w/o Entropy Reg., HTA, F2T	*lr = 0.001* *wd = 0.0001*	0.694***	0.434***	0.349***	0.586***
CAAC w/o Entropy Reg., HTA, T2F	*lr = 0.001* *wd = 0.001*	0.799***	0.657***	0.722***	0.692***
CAAC w/o Entropy Reg., HTA, Co-Attention	*lr = 0.01* *wd = 0.0*	0.814***	0.708***	0.748***	0.726***
CAAC w/o Entropy Reg., HTA	*lr = 0.1* *wd = 0.0001*	0.791***	0.683***	0.759***	0.716***
CAAC w/o Entropy Reg.	*lr = 0.0001* *wd = 0.0001*	0.794***	0.660***	0.742***	0.698***
CAAC (our method)	*lr = 0.0001* *wd = 0.0*	**0.809**	**0.748**	**0.759**	**0.753**

Note. (*** $p < 0.01$). Bold values indicate the best performance for each metric. Italic entries indicate tuned hyperparameter values. Pairwise t-tests on the held-out test set assess significance of differences between the full CAAC model and each ablation or baseline. Ablation variants disable one or more of these modules (e.g., "w/o Softmax" removes temporal aggregation, "w/o Co-Attention" removes cross-modal fusion, "w/o Entropy Reg." omits the regularization term). For the full model, $\alpha = 0.5, \beta = 1$; for ablations without entropy regularization, $\alpha = \beta = 0$.

Table E.2 Alternate Encoders

Model	Hyper-parameters	AUC	F1-Score	Precision	Recall
BERT + DINOv2	$lr = 0.001,\ wd = 0.01$ $\alpha = 0.5,\ \beta = 1$	0.712***	0.567***	0.709***	0.643***
Long^a + DINOv2	$lr = 0.001,\ wd = 0.01$ $\alpha = 0.5,\ \beta = 0.5$	0.737***	0.691***	0.705***	0.703***
Long^a + ViT	$lr = 0.01,\ wd = 0.001$ $\alpha = 0.5,\ \beta = 1$	0.757***	0.632***	0.713***	0.676***
CAAC† (our method)	$lr = 0.0001,\ wd = 0$ $\alpha = 0.5,\ \beta = 1$	**0.822**	**0.756**	**0.768**	**0.763**

Note. (*** $p < 0.01$). Bold values indicate the best performance for each metric. Italic entries indicate tuned hyperparameter values. Pairwise t-tests on the held-out test set are used to assess statistical significance of performance differences between CAAC and the comparison models. Long^a = Longformer. † Indicates the full CAAC configuration (BERT + ViT + Integrated Multimodal Fusion Module). For all alternate-encoder variants (e.g., replacing BERT $\rightarrow$ Longformer or ViT $\rightarrow$ DINOv2), the Integrated Multimodal Fusion Module and training settings remain identical.

Online Supplement F: Fusion Results

Table F.1 Performance of different encoder combinations under various fusion strategies, for Screen-Change (top) and Sentence-Based (bottom).

Encoder Combination	Early Fusion		Late Fusion		CAAC (Co-attention)	
	AUC	F1	AUC	F1	AUC	F1
Screen-Change						
Long^a + DINOv2	0.510***	0.434***	0.674***	0.573***	0.737***	0.691***
BERT + DINOv2	0.502***	0.434***	0.787***	0.728***	0.712***	0.567**
Long^a + ViT	0.501***	0.434***	0.749***	0.668***	0.757***	0.632***
BERT + ViT	0.517***	0.434***	0.801***	0.666***	**0.822**	**0.756**
Sentence-Based						
Long^a + DINOv2	0.575***	0.502***	0.674***	0.573***	0.753***	0.676***
BERT + DINOv2	0.801***	0.714***	0.787***	0.728***	0.774***	0.712***
Long^a + ViT	0.578***	0.499***	0.749***	0.668***	0.790***	0.722***
BERT + ViT	0.804***	0.718***	0.801***	0.666***	**0.809**	**0.748**

Note. (*** $p < 0.01$). Bold values indicate the best performance for each metric. Pairwise t-tests on the held-out test set assess statistical significance of performance differences between CAAC (our method) and the comparison models. Long^a = Longformer.

Table F.2 Aggregation Approaches at Video-level for Screen-Change (top) and Sentence-Based (bottom).

Encoder Combination	Attention		Mean		CAAC (Softmax)	
	AUC	F1	AUC	F1	AUC	F1
Screen-Change						
Long^a + DINOv2	0.681***	0.565***	0.762***	0.693***	0.737***	0.691***
BERT + DINOv2	0.720***	0.635***	0.757***	0.565***	0.712***	0.567***
Long^a + ViT	0.809***	0.724***	0.792***	0.702***	0.757***	0.632***
BERT + ViT	0.743***	0.451***	0.808***	0.698***	**0.822**	**0.756**
Sentence-Based						
Long^a + DINOv2	0.666***	0.472***	0.707***	0.539***	0.753***	0.676***
BERT + DINOv2	0.621***	0.500***	0.750***	0.664***	0.774***	0.712***
Long^a + ViT	0.789***	0.673***	0.774***	0.664***	0.790***	0.722***
BERT + ViT	0.699***	0.574***	0.791***	0.683***	**0.809**	**0.748**

Note. (*** $p < 0.01$). Bold values indicate the best performance for each metric. Pairwise t-tests on the held-out test set assess statistical significance of performance differences between CAAC (our method) and the comparison models. Long^a = Longformer.

Online Supplement G: Robustness Checks

Table G.1 COVID Videos

Encoders	Sentence-Based				Screen-Change			
	AUC	F1-Score	Precision	Recall	AUC	F1-Score	Precision	Recall
Long^a + ViT	0.680***	0.390***	0.302***	0.550***	0.643***	0.390***	0.302***	0.550***
CAAC (our method)	**0.794**	**0.513**	**0.680**	**0.600**	**0.800**	**0.493**	**0.768**	**0.600**

Note. (*** $p < 0.01$). Bold values indicate the best performance for each metric. Pairwise t-tests on the held-out test set are used to assess statistical significance of performance differences between CAAC (our method) and the comparison model, across fixed-format and screen change inputs. A total of 100 videos are taken where number of positive labels is 43 and negative labels is 57. Long^a = Longformer. Both the models include the same **ViT + Co-Attention + Softmax + Entropy Regularization** modules. Only the **text encoder** (BERT vs. Longformer) differs across configurations, allowing isolation of encoder-specific effects.

Table G.2 Performance across sample sizes

Sample Size	AUC	F1	Precision	Recall
1249	0.817**	0.754	0.758*	0.757**
1149	**0.822**	**0.756**	**0.768**	**0.763**
1049	0.823	0.746***	0.753***	0.751***
949	0.821	0.748***	0.767	0.758**
849	0.828***	0.747***	0.752***	0.746***
749	0.813***	0.719***	0.762	0.739***
649	0.823***	0.735***	0.743***	0.741***

Note. (* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$). Bold values indicate the best performance for each metric. Pairwise t-tests on the held-out test set are used to assess statistical significance of performance differences between our method original training sample size and the corresponding variations. The original training sample size is 1,149 videos.

Online Supplement H: Performance by Video Duration

Table H.1 Length-based performance split cutoff (less than 4 min)

Metric	Long + ViT			CAAC(Our Method)		
	$<=4mins$	$>4mins$	p-value	$<=4mins$	$>4mins$	p-value
AUC	0.791	0.717	0.170	0.828	0.794	0.481
F1-Score	0.711	0.561	**0.009**	0.800	0.704	**0.040**
Precision	0.779	0.652	**0.020**	0.803	0.727	0.093
Recall	0.754	0.598	**0.001**	0.808	0.710	**0.031**

Note. Independent two-sample t-tests (Welch) between videos ($<=4$ min; $n_1 = 187$) and videos (>4 min; $n_2 = 172$) within each model group; video sets are non-overlapping. 95% confidence intervals are reported where applicable. p-values are two-sided; bold denotes $p < 0.05$.