Seeing Beyond Words: Multimodal Aspect-Level Complaint Detection in E-commerce Videos

Rishikesh Devanathan, Apoorva Singh, A.S. Poornash, Sriparna Saha

ACM Multimedia 2024

Motivation

- Multimedia (video) **reviews increasingly shape purchasing decisions**
    - Understanding **user sentiments** and **pinpointing specific complaint aspects** within these videos can provide valuable insight to a company about their product
- Visuals and audio provide important details about complaints beyond just text
    - Cues like **reviewer tone, facial expression** and **design of the product** reveal a lot about the complaint and aspect

Aspect-Level Complaint Detection: Problem Formulation

VCD dataset is represented as $D = \{[T, V, A, a_k, c_k]_i \}_{i=1}^{N}$, where V and A denotes the review video and it's corresponding audio A. T is the review text, a_k and c_k denote the aspect and complaint category for the k^{th} clip spanning timestamp $t_k=\{t_{start,k}, t_{end,k}\}$ in the video V

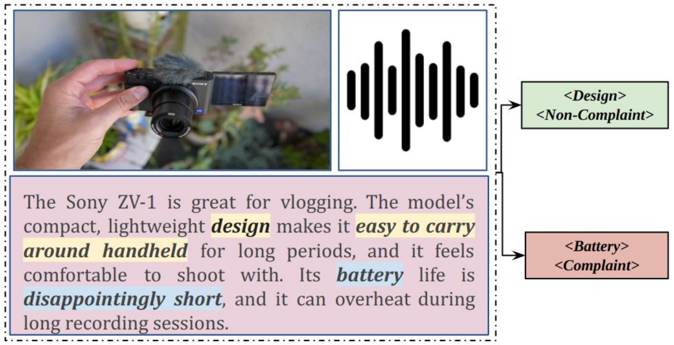
Review	Video, Audio	Aspect	Complaint
The device has Gorilla Glass 3 protection, ensuring good screen quality. Battery life appears to be commendable lasting around 3 days. But the front-facing camera, rated at 12 megapixels, offers only average performance.		Design	Non-Complaint
		Battery	Non-Complaint
		Camera	Complaint

Aspect-Level Complaint Detection: Problem Formulation

We define two tasks from the given dataset:

1. Aspect Complaint Classification (ACC): $c_i = M(T_i, V_i, A_i)$
2. Aspect Category Detection (ACD): $a_i = M(T_i, V_i, A_i)$

We propose a unified model M that predicts both the aspect (a_i) and complaint c_i in a multitask manner



The Sony ZV-1 is great for vlogging. The model's compact, lightweight ***design*** makes it ***easy to carry around handheld*** for long periods, and it feels comfortable to shoot with. Its ***battery*** life is ***disappointingly short***, and it can overheat during long recording sessions.

Challenges

Data Scarcity

Scarcity of datasets for complaint analysis in video reviews.

Automated Models

Lack of automated methods for analyzing diverse video content leveraging video, audio and text modalities

Contributions

Video Complaint Dataset (VCD)

- A novel resource aimed at advancing research in aspect-level complaint detection

Multimodal Aspect-Aware Complaint Analysis (MAACA)

- Multimodal (audio, video and text) pre-training for alignment
- Gated Fusion
- Moment retrieval augmentation

Video Complaint Dataset (VCD)

Related Work: Task and Dataset

- Complaint Cause Analysis: aiming to detect and extract the reasons behind Twitter complaints
  - Introduces interpretability dimension in complaint detection
- CESAMARD: collection of consumer feedback or reviews and images of products purchased from the e-commerce website Amazon

No video dataset for aspect-based complaint detection!

Video Complaint Dataset (VCD)

- 128 unique review videos from Youtube spanning different products (phones, laptops, cameras, etc.)
  - 443 total annotated instances
- Each annotation contains:
  1. Timestamp
  2. **Complaint Label:** (Yes/No)
  3. **Entity Label: (**Camera, Os, Design, Battery, Price, Speaker, Storage)
  4. Text (Transcript): Generated using Whisper or written manually
  5. Audio
  6. Video

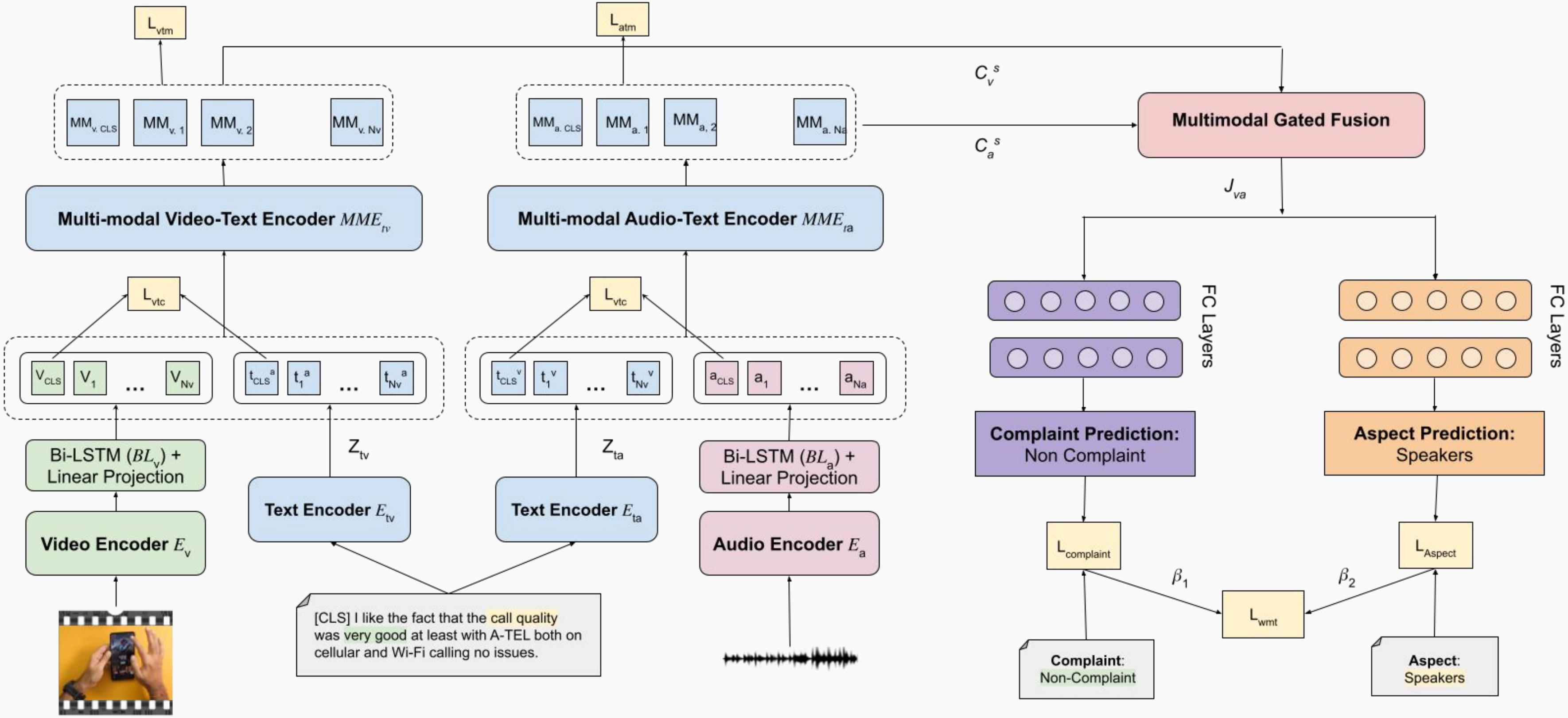
Review	Video	Aspect	Complaint
The device has Gorilla Glass 3 protection, ensuring good screen quality. Battery life appears to be commendable lasting around 3 days. But the front-facing camera, rated at 12 megapixels, offers only average performance.		Design	Non Complaint
		Battery	Non Complaint
		Camera	Complaint

Multimodal Aspect-Aware Complaint Analysis (MAACA)

Related Work: Models

- ToxVidLM: multi-modal features serve as the query and the embedding matrix of any LLM as the key and value
  - Embedding matrix is large ⇒ high memory, time consumption
  - Makes use of a information compression approach using convolution
- Video-Llava: makes use of LanguageBind which produces unified visual representations to the language feature space
  - Highlights the importance of alignment
- ALPRO: aligns the video and text encoders with contrastive and matching loss

We borrow ideas from these models to build MAACA!


L_{vtm}
L_{atm}
C_v^s
C_a^s
Multimodal Gated Fusion
J_{va}
$MM_{v. CLS}$ $MM_{v, 1}$ $MM_{v, 2}$ $MM_{v. Nv}$
$MM_{a. CLS}$ $MM_{a. 1}$ $MM_{a, 2}$ $MM_{a. Na}$
Multi-modal Video-Text Encoder MME_{tv}
Multi-modal Audio-Text Encoder MME_{ta}
L_{vtc}
L_{vtc}
V_{CLS} V_1 ... V_{Nv}
t_{CLS}^a t_1^a ... t_{Nv}^a
t_{CLS}^v t_1^v ... t_{Nv}^v
a_{CLS} a_1 ... a_{Na}
Z_{tv}
Z_{ta}
Bi-LSTM (BL_v) + Linear Projection
Bi-LSTM (BL_a) + Linear Projection
Video Encoder E_v
Text Encoder E_{tv}
Text Encoder E_{ta}
Audio Encoder E_a
FC Layers
FC Layers
Complaint Prediction: Non Complaint
Aspect Prediction: Speakers
$L_{complaint}$
L_{Aspect}
β_1
β_2
L_{wmt}
Complaint: Non-Complaint
Aspect: Speakers

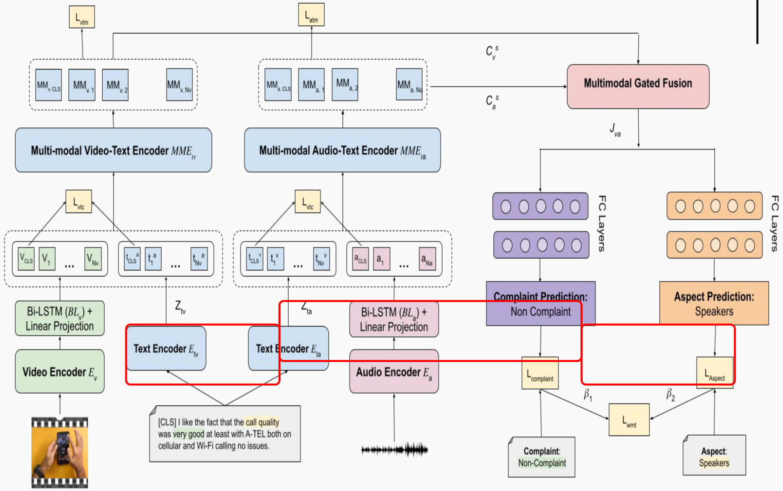
[CLS] I like the fact that the call quality was very good at least with A-TEL both on cellular and Wi-Fi calling no issues.

MAACA: Multimodal Aspect-Aware Complaint Analysis

MAACA: Encoders

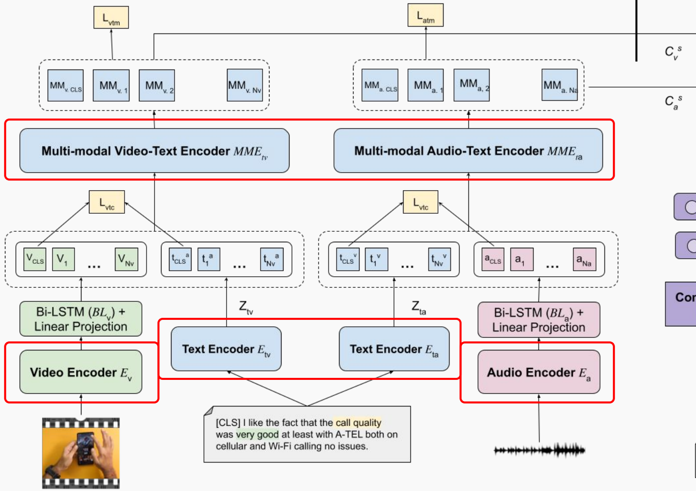
- Text encoders: two instances of the 6-layer **BERT-encoders** E_{tv} (video feature alignment) and E_{ta} (audio feature alignment)
  - Both the encoders output an embedding sequence of t = $\{t_{CLS}, t_1, \ldots t_{Nt}\}$ with $t_i \in R^D$

- Video encoder E_v: **TimeSFormer**
  - Clip is sampled uniformly into T=16 frames of size 224×224×3. Each frame is chunked into patches, flattened, and mapped to an embedding through **attention applied across time and spatial (patch) dimension**
  - Obtain video features Z'_v of dimension $v \in R^{N'vxDv}$ (N'$_v$ tends to be large)



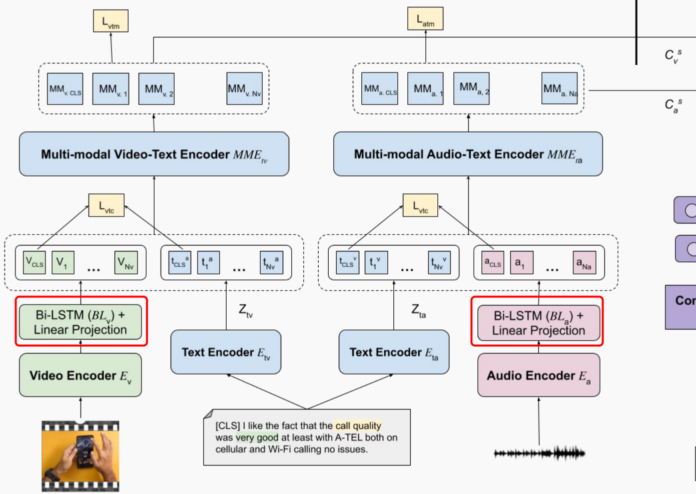
MAACA: Encoders

- Audio Encoders E_t: **whisper-small**
  - Audio corresponding to the text transcript is sampled at the rate of 16000 Hz, split into 30 second intervals and passed to whisper-small
  - Passed into an encoder consisting of a series of self-attention blocks
  - Output of the audio encoder has a dimension of $a \in R^{N'axDa}$ (N'$_a$ tends to be large)



MAACA: Information Compression

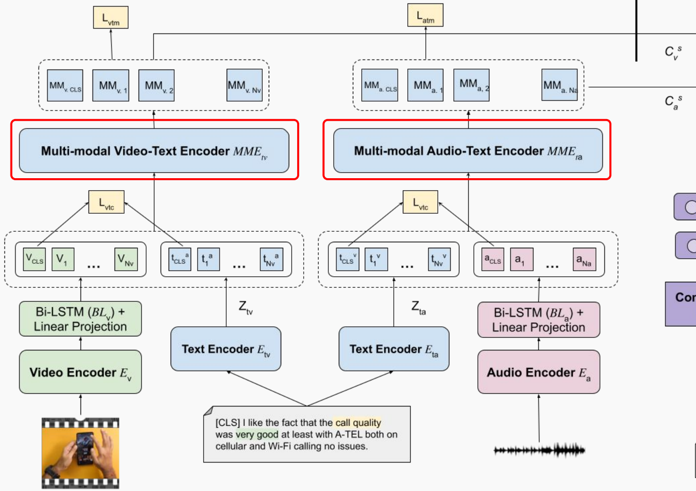
- N'_a and N'_v tend to be large compared to N_t = 128
- Time complexity of cross-attention (next step) is proportional to sequence length and dimension of the embedding ($O(N^2D)$)
  - Thus we can save time and memory by condensing audio and video sequence
- Original encoder outputs are fed into a Bi-LSTM
- Output sequence consists of the last N_m timesteps
  - $N_m < N'_m$, $m \in \{a, v\}$
  - Rationale: last N_m timesteps are significant as they are encapsulate the cumulative information of the preceding and successive timesteps
- The output embedding is projected to common dimension D (corresponding to the text embedding)



L_{vtm}
L_{atm}
C_v^s
$MM_{v.\,CLS}$ $MM_{v.\,1}$ $MM_{v.\,2}$ $MM_{v.\,Nv}$
$MM_{a.\,CLS}$ $MM_{a.\,1}$ $MM_{a,\,2}$ $MM_{a.\,Na}$
C_a^s
Multi-modal Video-Text Encoder MME_{tv}
Multi-modal Audio-Text Encoder MME_{ta}
L_{vtc}
L_{vtc}
V_{CLS} V_1 ... V_{Nv}
t_{CLS}^a t_1^a ... t_{Nv}^a
t_{CLS}^v t_1^v ... t_{Nv}^v
a_{CLS} a_1 ... a_{Na}
Bi-LSTM (BL_v) + Linear Projection
Z_{tv}
Z_{ta}
Bi-LSTM (BL_a) + Linear Projection
Video Encoder E_v
Text Encoder E_{tv}
Text Encoder E_{ta}
Audio Encoder E_a
[CLS] I like the fact that the call quality was very good at least with A-TEL both on cellular and Wi-Fi calling no issues.
Cor

MAACA: Multimodal Encoder

- Concatenate the outputs of the text encoders with video encoder and audio encoder to produce Y_{tv} and Y_{ta} which are the inputs to the MME
- Two instances of the 6-layer **BERT-encoders** MME$_{tv}$ (video feature alignment) and MME$_{ta}$ (audio feature alignment)
  - ○ Consists of a series of self-attention blocks
- Outputs sequence of multimodal (MM) tokens $MM_{m,CLS}$, $MM_{m,1}$, . . . $MM_{m,Nm+Nt}$ with each token $MM_i \in R^D$, where $m \in \{a, v\}$

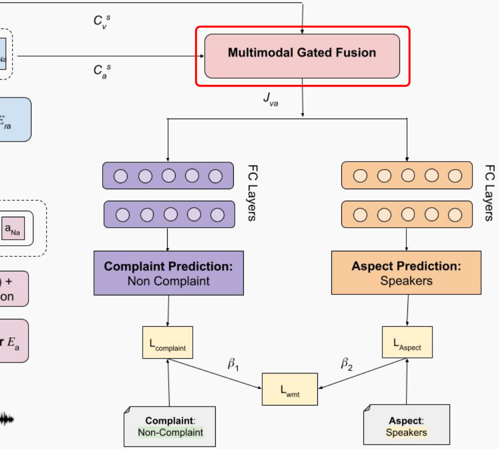


L_{vtm}
L_{atm}
C_v^s
C_a^s
$MM_{v.CLS}$ $MM_{v.1}$ $MM_{v.2}$ $MM_{v.Nv}$
$MM_{a.CLS}$ $MM_{a.1}$ $MM_{a,2}$ $MM_{a.Na}$
Multi-modal Video-Text Encoder MME_{tv}
Multi-modal Audio-Text Encoder MME_{ta}
L_{vtc}
L_{vtc}
V_{CLS} V_1 ... V_{Nv}
t_{CLS}^a t_1^a ... t_{Nv}^a
t_{CLS}^v t_1^v ... t_{Nv}^v
a_{CLS} a_1 ... a_{Na}
Z_{tv}
Z_{ta}
Bi-LSTM (BL_v) + Linear Projection
Bi-LSTM (BL_a) + Linear Projection
Video Encoder E_v
Text Encoder E_{tv}
Text Encoder E_{ta}
Audio Encoder E_a
Cor

[CLS] I like the fact that the call quality was very good at least with A-TEL both on cellular and Wi-Fi calling no issues.

MAACA: Multimodal Gated Fusion

- How to combine the text-aligned video (C^s_v) and text-aligned audio (C^s_a) tokens?
- Simple weighting (same weight for all positions) vs Gating (varying weights)



1. **Calculate the gate α from the input tokens**

   $$\alpha = \sigma(P_v C^s_v + P_a C^s_a + b_g)$$

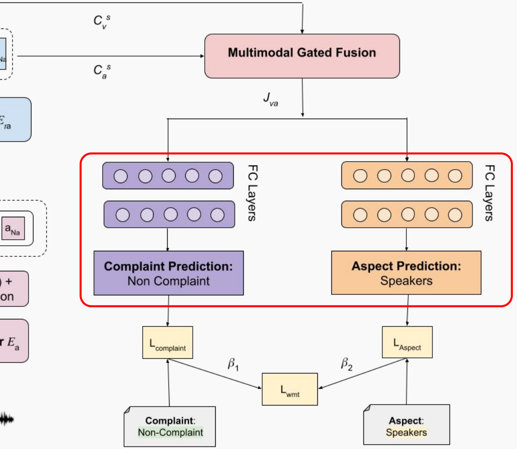
   P_v and P_a represent weight matrices for the visual and acoustic modalities, while b_g denotes scalar bias and σ is the sigmoid activation function

2. **Final joint representation J_{va} is represented as**

   $$J_{va} = \alpha\, C^s_a + (1 - \alpha) C^s_v$$

MAACA: Multitask Heads

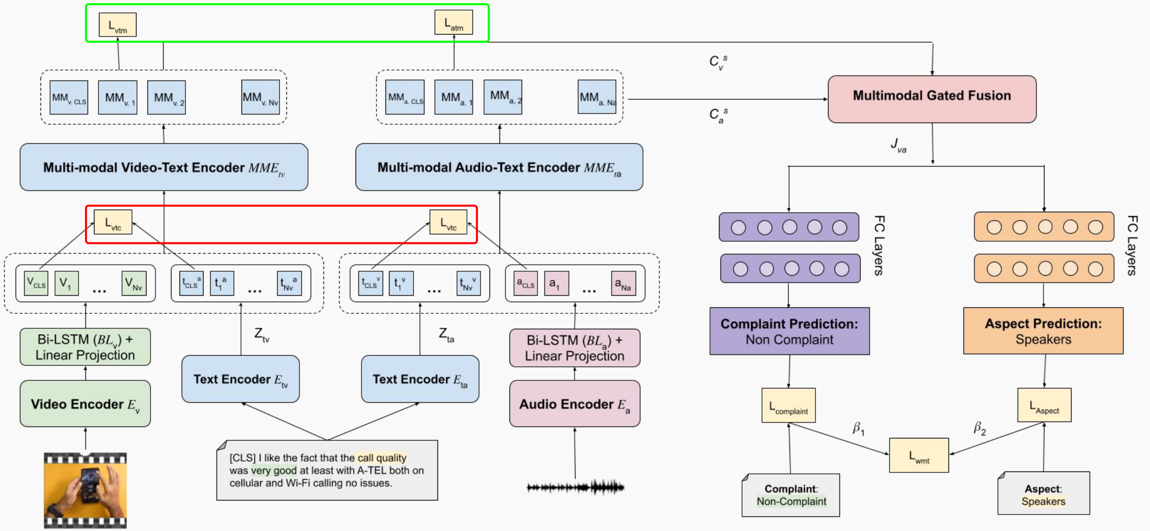
- Two task-specific fully connected heads for simultaneously predicting the complaint and aspect classes.



- Comprises two linear layers with a softmax function finally producing the class probabilities of appropriate shape
  - 2 for the complaint/non-complaint classification task (ACC)
  - 7 for the aspect identification task (ACD)

Pretraining

- Can we directly concatenate v_i and t_i^v or a_i and t_i^a to create the input for the multimodal encoder?
- Is the multimodal encoder trained to produce coherent output?



Pretraining: Contrastive Loss (VTC, ATC)

- Objective: align the **unimodal video and audio representations** with their text counterparts before passing to the multimodal encoder

1. Similarity between i^{th} audio/video feature with the j^{th} text feature is

$$s(Y_{m,i}, Z_{tm,j}) = m_{CLS,i} \cdot t_{CLS,j}$$

2. Calculate the negative log likelihood terms using the similarity function (B = batch size)

$$\mathcal{L}_{\mathrm{m2t}} = -\log \frac{\exp\left(s\left(Y_{m,i}, Z_{tm,i}\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(s\left(Y_{m,i}, Z_{tm,i}\right)/\tau\right)}$$

$$\mathcal{L}_{\mathrm{t2m}} = -\log \frac{\exp\left(s\left(Z_{tm,i}, Y_{m,i}\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(s\left(Z_{tm,i}, Y_{m,i}\right)/\tau\right)}$$

3. Final loss is the mean of the two terms above:

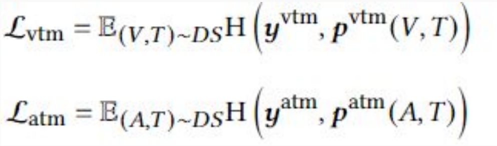
$$\mathcal{L}_{\mathrm{vtc}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{v2t}} + \mathcal{L}_{\mathrm{t2v}}\right)$$

$$\mathcal{L}_{\mathrm{atc}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{a2t}} + \mathcal{L}_{\mathrm{t2a}}\right)$$

Pretraining: Matching Loss (VTM, ATM)

- Objective: aims to **align the multimodal encoder** by learning to distinguish the positive video/text and audio/text pairs from the negative ones.

1. Multimodal encoder outputs MM tokens $\{MM_{m,CLS}, MM_{m,1}, \ldots MM_{m,Nm+Nt}\}$
2. Mine hard negative pairs <T, V> and <T, A> from the batch of <T, V, A> by using the similarity function described earlier.
3. Pass $MM_{m,CLS}$ to a fully connected layer to predict $\boldsymbol{p}^{\text{vtm}}$ representing probability that each <T,V> or <T,A> pairs represent the same instance
4. Matching loss is calculated as the cross entropy loss between the one hot vector $\boldsymbol{y}^{\text{vtm}}$ and the probability distribution $\boldsymbol{p}^{\text{vtm}}$



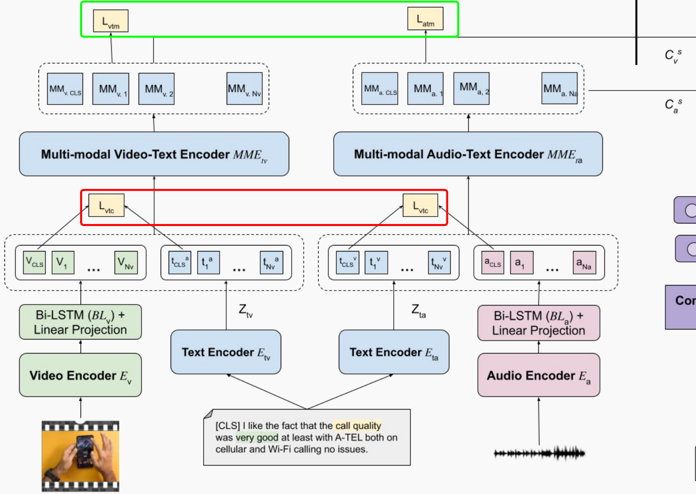
$$\mathcal{L}_{\text{vtm}} = \mathbb{E}_{(V,T)\sim DS}\mathrm{H}\left(\boldsymbol{y}^{\text{vtm}}, \boldsymbol{p}^{\text{vtm}}(V,T)\right)$$

$$\mathcal{L}_{\text{atm}} = \mathbb{E}_{(A,T)\sim DS}\mathrm{H}\left(\boldsymbol{y}^{\text{atm}}, \boldsymbol{p}^{\text{atm}}(A,T)\right)$$

Cumulative Pretraining Loss

- Independently align the video and audio with the text modality with
  - Contrastive loss: aligns the unimodal encoders
  - Matching loss: aligns the multimodal encoders
- Pre-trained for 4 epochs with batch size of 8

$$L_{pre-training} = L_{vtc} + L_{atc} + L_{vtm} + L_{atm}$$

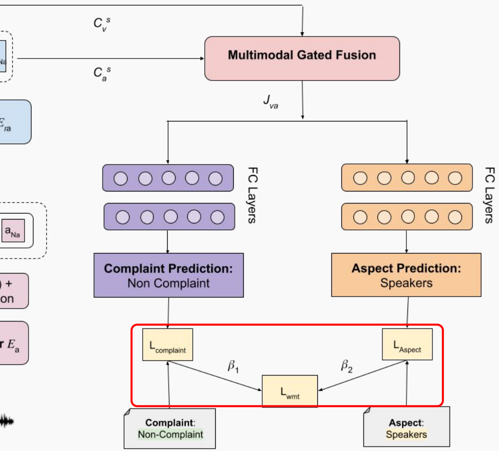


L_{vtm}
L_{atm}
$MM_{v, CLS}$
$MM_{v, 1}$
$MM_{v, 2}$
$MM_{v, Nv}$
$MM_{a, CLS}$
$MM_{a, 1}$
$MM_{a, 2}$
$MM_{a, Na}$
C_v^s
C_a^s
Multi-modal Video-Text Encoder MME_{tv}
Multi-modal Audio-Text Encoder MME_{ta}
L_{vtc}
L_{vtc}
V_{CLS} V_1 ... V_{Nv}
t_{CLS}^a t_1^a ... t_{Nv}^a
t_{CLS}^v t_1^v ... t_{Nv}^v
a_{CLS} a_1 ... a_{Na}
Z_{tv}
Z_{ta}
Bi-LSTM (BL_v) + Linear Projection
Bi-LSTM (BL_a) + Linear Projection
Video Encoder E_v
Text Encoder E_{tv}
Text Encoder E_{ta}
Audio Encoder E_a
Cor

[CLS] I like the fact that the call quality was very good at least with A-TEL both on cellular and Wi-Fi calling no issues.

Fine-tuning Loss

- With the aligned encoders, we fine-tune the model with the objective of complaint and aspect classification



- The multitask head produce class probabilities for aspect and complaint classes
- Fine-tuning loss is expressed as the weighted sum of cross-entropy loss for aspect and complaint prediction:

$$Loss_{wmt} = \sum_{k=1}^{M} \beta_k Loss_k$$

- Trained for 10 epochs with early stopping

Preprocessing: Moment Retrieval

- Review videos often include extended introductions, outros, or cutaway shots of the reviewer's face
- Use moment retrieval to focus on relevant segments in the clip
- Make use of CG-DETR model, which achieves state-of-the-art results in QVHighlights dataset

1. Each clip is passed into the CG-DETR with a set of simple prompts denoting all the product categories (Eg "Phone", "Laptop", "Camera", etc.)
2. Time frame corresponding to the highest saliency score across all the prompts is taken.
3. The video is clipped to that time frame

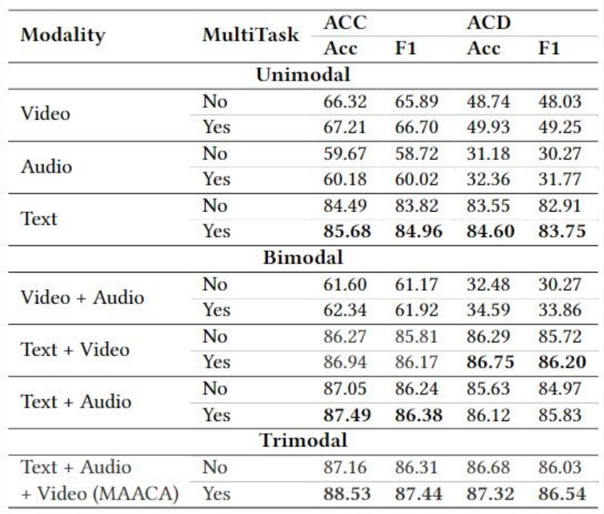
Results: Unimodal Baselines

Unimodal Model:

1. Derived final representations using modality-specific encoders. (Z_{ta}/Z_{tv}, Z_a, Z_v)
2. Subjected to a CLS pooling operation, resulting in the extraction of overall semantic information contained within each modality (m_{CLS})
3. Pooled output is passed to the prediction head and trained in multitask/single task manner.

Insights:

Text (transcripts), contains the richest information for both the complaint (ACC) and aspect (ACD) tasks (this is why we align audio and video with text)



Modality	MultiTask	ACC		ACD	
		Acc	F1	Acc	F1
		Unimodal			
Video	No	66.32	65.89	48.74	48.03
	Yes	67.21	66.70	49.93	49.25
Audio	No	59.67	58.72	31.18	30.27
	Yes	60.18	60.02	32.36	31.77
Text	No	84.49	83.82	83.55	82.91
	Yes	**85.68**	**84.96**	**84.60**	**83.75**
		Bimodal			
Video + Audio	No	61.60	61.17	32.48	30.27
	Yes	62.34	61.92	34.59	33.86
Text + Video	No	86.27	85.81	86.29	85.72
	Yes	86.94	86.17	**86.75**	**86.20**
Text + Audio	No	87.05	86.24	85.63	84.97
	Yes	**87.49**	**86.38**	86.12	85.83
		Trimodal			
Text + Audio + Video (MAACA)	No	87.16	86.31	86.68	86.03
	Yes	**88.53**	**87.44**	**87.32**	**86.54**

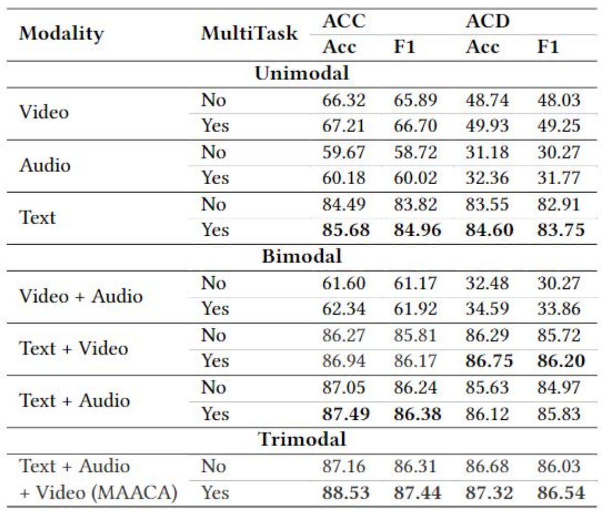
Results: Bimodal Baselines

Bimodal Models

1. Concatenated unimodal representation is passed to the multimodal encoder to obtain the multimodal representation C^{s}_{m}
2. Pooling operation is applied on the CLS token $MM_{m,CLS}$ of the multimodal representation
3. Pooled output is used to predict aspect and complaint

Insights

- Models incorporating text consistently outperform those relying solely on video and audio inputs
- **Audio cues are crucial for complaint detection** and **visual information is crucial to aspect identification**!

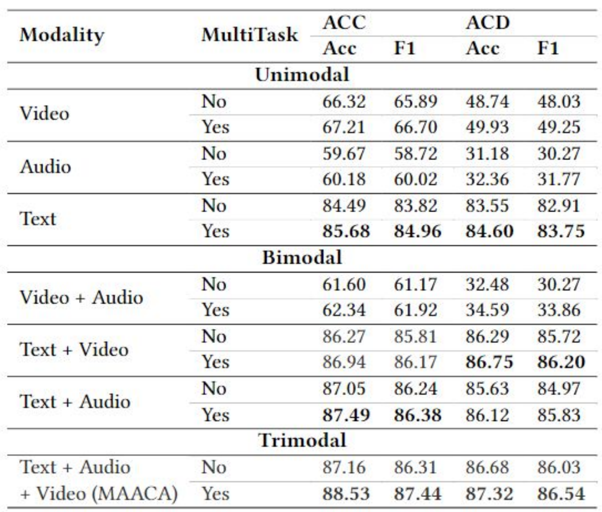


Modality	MultiTask	ACC		ACD	
		Acc	F1	Acc	F1
		Unimodal			
Video	No	66.32	65.89	48.74	48.03
	Yes	67.21	66.70	49.93	49.25
Audio	No	59.67	58.72	31.18	30.27
	Yes	60.18	60.02	32.36	31.77
Text	No	84.49	83.82	83.55	82.91
	Yes	**85.68**	**84.96**	**84.60**	**83.75**
		Bimodal			
Video + Audio	No	61.60	61.17	32.48	30.27
	Yes	62.34	61.92	34.59	33.86
Text + Video	No	86.27	85.81	86.29	85.72
	Yes	86.94	86.17	**86.75**	**86.20**
Text + Audio	No	87.05	86.24	85.63	84.97
	Yes	**87.49**	**86.38**	86.12	85.83
		Trimodal			
Text + Audio + Video (MAACA)	No	87.16	86.31	86.68	86.03
	Yes	88.53	87.44	87.32	86.54

Results: Trimodal (MAACA)

- Trimodal: Leverage the **complementary strengths** of each modality
- Beats the ACC F1 score of
  - Text-only model by 2.48%
  - Text+Audio model by 1.06%
- Beats the ACD F1 score of
  - Text-only model by 2.79%
  - Text+Video model by 0.34%

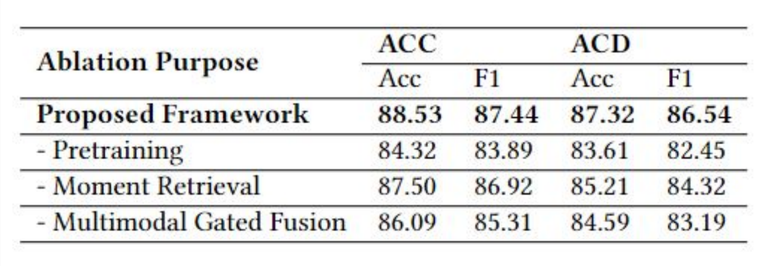
- Multi-task models outperform single-task counterparts for most experiments.
  - ACC and ACD are complementary tasks



Modality	**MultiTask**	**ACC**		**ACD**	
		Acc	**F1**	**Acc**	**F1**
		Unimodal			
Video	No	66.32	65.89	48.74	48.03
	Yes	67.21	66.70	49.93	49.25
Audio	No	59.67	58.72	31.18	30.27
	Yes	60.18	60.02	32.36	31.77
Text	No	84.49	83.82	83.55	82.91
	Yes	**85.68**	**84.96**	**84.60**	**83.75**
		Bimodal			
Video + Audio	No	61.60	61.17	32.48	30.27
	Yes	62.34	61.92	34.59	33.86
Text + Video	No	86.27	85.81	86.29	85.72
	Yes	86.94	86.17	**86.75**	**86.20**
Text + Audio	No	87.05	86.24	85.63	84.97
	Yes	**87.49**	**86.38**	86.12	85.83
		Trimodal			
Text + Audio + Video (MAACA)	No	87.16	86.31	86.68	86.03
	Yes	**88.53**	**87.44**	**87.32**	**86.54**

Results: Ablation Studies

- **Pre-training plays the biggest role:** increase of 3.55% in complaint and 4.09% in aspect F1
    - Aligned encoders result in better learning after downstream fine-tuning
- **Gated attention plays the second biggest role**
    - Instead of gating C^s_v and C^s_a, they are concatenated and pooled. The pooled output is used for classification
- **Moment retrieval plays the third biggest role**
    - When moment retrieved clips are used as video input, it increases both ACC and ACD F1-scores by 0.52% and 2.22%



Ablation Purpose	ACC		ACD	
	Acc	F1	Acc	F1
Proposed Framework	**88.53**	**87.44**	**87.32**	**86.54**
- Pretraining	84.32	83.89	83.61	82.45
- Moment Retrieval	87.50	86.92	85.21	84.32
- Multimodal Gated Fusion	86.09	85.31	84.59	83.19

Thank You