

# Synthesizing Sentiment-Controlled Feedback For Multimodal Text and Image Data

Puneet Kumar <sup>†</sup> , Member, IEEE, Sarthak Malik <sup>†</sup> , Balasubramanian Raman , Senior Member, IEEE, and Xiaobai Li <sup>\*</sup> , Senior Member, IEEE

**Abstract**—The ability to generate sentiment-controlled feedback in response to multimodal inputs—comprising both text and images—addresses a critical gap in human-computer interaction by enabling systems to provide empathetic, accurate, and engaging responses. This capability has profound applications in healthcare, marketing, and education. To this end, we construct a large-scale Controllable Multimodal Feedback Synthesis (CMFeed) dataset and propose a controllable feedback synthesis system. The proposed system includes an encoder, decoder, and controllability block for textual and visual inputs. It extracts textual and visual features using a transformer and Faster R-CNN networks and combines them to generate feedback. The CMFeed dataset encompasses images, text, reactions to the post, human comments with relevance scores, and reactions to the comments. The reactions to the post and comments are utilized to train the proposed model to produce feedback with a particular (positive or negative) sentiment. A sentiment classification accuracy of 77.23% has been achieved, 18.82% higher than the accuracy without using the controllability. Moreover, the system incorporates a similarity module for assessing feedback relevance through rank-based metrics. It implements an interpretability technique to analyze the contribution of textual and visual features during the generation of uncontrolled and controlled feedback.

**Index Terms**—Natural Language Generation, Controllability, Interpretability, Multimodal Analysis, Affective Computing.

## 1 INTRODUCTION

THE process of multimodal feedback synthesis involves generating responses to multimodal inputs in a way that mirrors human feedback [1]. Controlling sentiments in feedback, a capability inherent to humans, remains a challenge for machines [2]. The ability to control sentiments in feedback synthesis has wide-ranging applications. For instance, it can lead to more empathetic patient care in healthcare, more accurate consumer insights in marketing, and more engaging educational material development. Such systems can predict the mental states of patients, assess user responses to products, analyze social behaviors, evaluate multimodal educational content, and predict user engagement in advertisements [3], [4], [5]. Crucially, the controllability aspect of these systems enables them to adopt different roles in human-computer interaction (HCI), such as responding in varied styles to suit the audience’s needs – from encouraging the stressed or depressed, acting as a strict coach, to being delightful or bubbly in lighter contexts. The ability to control sentiments in feedback enriches HCI, making it more personalized.

This paper introduces the task of controllable feedback synthesis for input images and text. It proposes a large-scale Controllable Multimodal Feedback Synthesis (CMFeed) dataset for this purpose, created by crawling Facebook news articles. A feedback synthesis system has been developed that generates sentiment-controlled feedback with positive or negative sentiments. It employs two networks for textual and visual modalities, each comprising an

encoder, a decoder, and a control layer. The encoders use a text transformer [6] and a Faster Region-Based Convolutional Neural Networks (R-CNN) model [7] to extract features, which the decoder combines for feedback generation. The control layer, placed after the decoder, selectively activates or deactivates neurons corresponding to the positive and negative sentiments. The system also includes a similarity module, which ensures that generated feedback closely aligns with the input’s context, enhancing the relevance and accuracy of the response. Furthermore, the contribution of various textual and visual features toward generating uncontrolled and controlled feedback has been analyzed using the K-Average Additive exPlanation (KAAP) interpretability technique.

The proposed task of controllable feedback synthesis differs from existing related tasks. Unlike multimodal summarization [8], which condenses information, or dialog generation [9], which focuses on conversation flow, it requires creating contextually apt and sentiment-aligned feedback. Importantly, our work focuses on generating ‘opinions’, rather than reusing ‘knowledge’ or ‘facts’. Our approach uniquely allows for the generation of controlled opinions, which can vary just as human opinions do. Visual question answering [10] and sentiment-aware conversation [11] involve extracting or inferring sentiments but do not generate sentiment-influenced feedback. Similarly, sentiment-controlled text generation [12], though sentiment-centric, needs the multimodal dimension of our approach. These tasks are empirically found incapable of synthesizing meaningful sentiment-controlled feedback [1], [13]. Furthermore, although Generative Pre-Trained Transformer (GPT) and Large Language Models (LLMs) [14] excel in summarization and question answering through knowledge retrieval, our approach fundamentally differs. Instead of relying on pre-existing data, our system interprets multimodal inputs to generate novel, sentiment-controlled feedback that mirrors human interaction. This highlights the distinctive contribution of our approach compared to GPT and other LLMs, which primarily leverage existing knowledge.

<sup>\*</sup> Corresponding Author.

<sup>†</sup> Contributed Equally.

P. Kumar is with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Email: puneet.kumar@oulu.fi

S. Malik and B. Raman are with Indian Institute of Technology Roorkee, India. Email: sarthak\_m@mt.iitr.ac.in and bala@cs.iitr.ac.in

X. Li is with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China, and the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Email: xiaobai.li@oulu.fi

The proposed system achieved sentiment classification accuracy of 77.23%, which is 18.82% higher than not using the controllability module. We analyzed the generated feedback's relevance and ranks against the human generated comments and also carried out human evaluations to assess the sentiment alignment. The major contributions of this paper are summarized as follows:

- The new task has been defined to generate sentiment-controlled feedback towards multimodal input containing images and text similar to how humans do.
- The CMFeed dataset has been constructed for multimodal feedback synthesis, encompassing text, images, corresponding comments, number of likes, shares, and sentiment class.
- A feedback synthesis system capable of generating sentiment-controlled feedback has been developed in which textual and visual features are extracted using a text transformer and a Faster R-CNN model and combined by the decoder for feedback generation.
- A novel controllability module has been introduced, enabling sentiment regulation in the generated feedback. It selectively activates or deactivates neurons to ensure alignment of the feedback with the desired sentiment.

The remainder of the paper is organized as follows: the existing related works are surveyed in Section 2. Section 3 describes the dataset construction process and the proposed system. The experiments and results have been discussed in Section 5 and Section 5.4 respectively, and the conclusions are drawn in Section 6.

## 2 RELATED WORKS

This section explores the state-of-the-art in the tasks related to controllable feedback synthesis and highlights research gaps.

### 2.1 Multimodal Summarization

Multimodal summarization combines the inputs from various modalities to produce comprehensive summaries [15]. In this direction, Chen and Zhuge [16] used recurrent networks to align text with corresponding images. In another work, Zhu et al. [17] focused on extractive summarization, selecting relevant images and texts using a pointer generator network. Their work was extended by Zellers et al. [18] for video summarization, using self-attention mechanisms to choose relevant video frames based on semantic content. Shang et al. [19] emphasized timing in video summarization with time-aware transformers. In other studies, Zhao et al. [20] explored audiovisual video summarization including sensory input and Xie et al. [21] used visual aesthetics while generating multimodal summaries.

Despite the advancements mentioned above, multimodal summarization faces the challenge of modality bias [22], [23], and it has not been explored for affect synthesis. Moreover, multimodal summarization approaches do not use human comments in model training. In contrast, the proposed system is trained on human-generated comments alongside text and image inputs to generate sentiment-controlled feedback. Furthermore, unlike multimodal summarization, which condenses the given information, controllable feedback synthesis produces feedback that aligns with sentiments and fits the context.

### 2.2 Visual Question Answering

Visual Question Answering (VQA) combines visual perception with interactive question answering [10]. Chen et al. [24] explored controlled generation versus standard answering methods in VQA,

while Wang et al. [25] focused on fact-based control to improve response accuracy. Cascante et al. [26] introduced an innovative approach using synthetic data, including 3D and physics simulations, to expand VQA's scope and enhance model adaptability. Furthermore, Jin et al. [27] proposed a text-based VQA system that explored the correlation between the text in the images and the visual scene. In another work, Guo et al. [28] developed a universal quaternion hypergraph network for advanced multimodal video question answering.

Contributions like Wu et al.'s [29] memory-aware control in community question answering and Lehmann et al.'s [30] use of language models for knowledge graph question answering illustrate VQA's diverse applications. These developments in VQA, however, stand in contrast to the proposed task of sentiment-controlled feedback synthesis. While VQA focuses on generating accurate responses to questions, sentiment-controlled feedback synthesis is centered on creating responses to multimodal stimuli (images and text) that are contextually relevant.

### 2.3 Dialogue Generation

In dialogue generation, particularly in visual dialogue (VisDial), computational models are designed to engage in dialogues about the visual content present in images. In this direction, Kang et al. [31] developed the Dual Attention Network (DAN) that utilized the understanding of visual references in dialogues. This approach was further advanced by Jiang et al. [32], who implemented a region-based graph attention network for enhanced image-related questioning. In another work, Xu et al. [33] and Zhao et al. [34] worked on modeling the conversations utilizing conditional variational autoencoders for diverse dialogue scenarios.

Recent developments in VisDial include the work of Chen et al. [35], who applied contrastive learning for better cross-modal comprehension in visual dialogue. Contributions from Wang et al. [36] and Kang et al. [37] involved enhanced dialogue generation through generative self-training. Liu et al. [38] and Liu et al. [39] also contributed to closed-loop reasoning and counterfactual visual dialogue for unbiased knowledge training. Despite aforementioned developments, existing VisDial methods do not generate sentiment-controlled feedback. The proposed task stands apart as it synthesizes feedback that is not only contextually relevant but is also tailored to the sentiment of the conversation.

### 2.4 Sentiment-Aware Conversation Agents

In the context of sentiment-aware conversational agents, Das et al. [11] laid the groundwork by using multiple generators and a multi-class discriminator to create text with diverse sentiment tones. Building upon this, Shi et al. [40] introduced a framework leveraging user sentiments to guide dialogue actions. In another work, Kong et al. [41] explored sentiment-constrained dialogue generation using generative adversarial networks. In recent developments, Firdaus et al. [42] presented a system that integrates sentiments into conversational agents. Their SEPRG model complemented it [43], which focuses on sentiment-aware, sentiment-controlled personalized response generation.

In other works, Hu et al. [44] developed a speech sentiment-aware agent for empathetic responses, and Saha et al. [45] integrated sentiment-aware mechanisms into conversation generation. Unlike the aforementioned works that primarily focus on recognizing the users' sentiments and using the same to influence dialogue actions, the proposed system generates (as opposed to recognizing) the feedbacks with the desired sentiment tone.

## 2.5 Sentiment-Controlled Text Generation

Recent natural text generation research trends include sentiment-controlled text generation. In this context, Zhou et al. [12] introduced the Emotional Chatting Machine (ECM), employing a seq2seq framework with internal and external memories for managing emotional states. In another work, Wang and Wan [46] generated sentimental texts using mixture adversarial networks. Huang et al. [47] contributed to automatic dialogue generation with a focus on expressed emotions. Building on these, Zhong et al. [48] introduced an affective attention mechanism with a weighted cross-entropy loss for generating affective dialogue.

The sentiment-controlled text generation methods do not use human comments to train their models. On the contrary, the proposed system captures the wider sentiment context from multimodal (text and images) inputs. Moreover, it is trained on human-generated comments aside from text and images to learn the contextual diversity from the comments. Furthermore, our work generates ‘opinions’ in contrast to generation of ‘knowledge’ or ‘facts’ by the above mentioned tasks. Unlike facts, the opinions can be controlled in a similar way how humans do. This area has not been addressed by the tasks mentioned earlier, while our work actively contributes towards it.

## 3 PROPOSED SYSTEM

This section describes the formulation of controllable feedback synthesis task and the architecture of the proposed system.

### 3.1 Task Formulation

Given an environment  $E = \{T, I_1, I_2, \dots, I_n\}$ , where  $T$  denotes input text and  $I_1, I_2, \dots, I_n$  denote  $n$  input images. Each image is comprised of  $m$  objects  $o_1, o_2, \dots, o_m$ , while the text  $T$  is made up of a dictionary of  $k$  words  $w_1, w_2, \dots, w_k$ . The task is to generate a feedback towards environment  $E$  under a sentiment-controlled manner where ‘sentiment-controlled’ implies the feedback aligns with a specified sentiment  $S$ , with  $S$  being either 0 for a negative sentiment or 1 for a positive sentiment.

### 3.2 Proposed System

The proposed system, as illustrated in Figure 1, employs two networks to process textual and visual data. Each network consists of an encoder, a decoder, and a control layer. The input data is processed by the encoder-decoder framework, followed by the control layer in both networks. The encoders, utilizing a text transformer [6] and a Faster R-CNN model [7], extract textual and visual features from the data. The decoder then uses these features to generate the feedback. The proposed system also includes a similarity module that evaluates the contextual similarity of the generated feedback with the human comments and an interpretability module that analyses the influence of different features on feedback generation. The proposed system implements both visual and textual attention to optimize information extraction and enhance overall performance. In the context of visual feature extraction, a pre-trained VGG network for global features’ extraction and Faster R-CNN for local features’ extraction have been used.

#### 3.2.1 Textual Encoder

The textual encoder is based on the transformer [6] and comprises global encoding and textual attention mechanisms. The global

encoding uses a convolution-gated unit to enhance textual representation, aiming to minimize repetition and maintain semantic relevance. Following the transformer model principles outlined in [6], the textual attention component includes multi-headed self-attention with a self-attention layer and a feed-forward layer. These components are further augmented with positional embedding to capture token positioning, and the process is finalized with normalization to yield the textual context vector  $z_t^*$ . The feed-forward network (FFN) within this setup features an input and output layer of dimension 512 and a hidden layer of dimension 2048. The FFN’s output for a particular input  $x$  is determined as specified in Eq.1.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (1)$$

where,  $b_1$ ,  $b_2$ ,  $W_1$ , and  $W_2$  represent the bias terms and weight matrices, respectively. In the self-attention mechanism, the query, key, and value weight matrices are initially randomized in the encoder and updated during training.

#### 3.2.2 Visual Encoder

The top three images from each sample are inputted into a visual encoder, using blank images for samples with fewer than three images. Their extracted features are concatenated into a visual context vector  $z_i^*$ . Feature extraction employs a pre-trained Faster R-CNN model [7]. As per Eq. 2, CNN layers output feature maps for the Region Proposal Network (RPN) to generate anchor boxes with binary scores based on Intersection Over Union (IoU) values [49]. These anchors undergo classification and regression, leading to classified boxes. A total of 1601 classes are assigned to these boxes, with their features combined to form a global feature vector. The Faster R-CNN has been chosen for object detection due to its efficiency and accuracy in handling small and varied objects, especially for non-real-time settings. The decision to use the top three images in the visual encoder is a considered trade-off between not losing significant visual content and avoiding the addition of blank images in posts with fewer images as most of the posts have at least 3 images.

$$Objectiveness Score = \begin{cases} Positive; & IoU > 0.7 \\ Negative; & IoU < 0.3 \\ No score; & 0.3 < IoU < 0.7 \end{cases} \quad (2)$$

#### 3.2.3 Attention

The attention mechanism in both the encoder and decoder operates on three vectors:  $Q$  (query),  $K$  (key), and  $V$  (value). The output of the self-attention layer, denoted as  $z_i$ , is computed by multiplying the  $i^{th}$  input vector of the encoder with the respective weight matrices  $W(Q)$ ,  $W(K)$ , and  $W(V)$ . This computation yields the attention head matrix  $z$ , as detailed in Eq. 3, whose dimensionality is equivalent to the length of the input sequence.

$$z = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \quad (3)$$

where  $Q$ ,  $K$ , and  $V$  represent matrices that contain all queries, keys, and values, respectively, with  $d_k$  as the scaling factor and  $K^T$  denoting the transpose of  $K$ . To achieve a comprehensive subspace representation, the mechanism computes multiple attention heads using distinct sets of Query, Key, and Value matrices. The queries, keys, and values undergo projection  $head$  times, resulting in multiple attention heads  $h_1, h_2, \dots, h_{head}$ , where  $head$  signifies

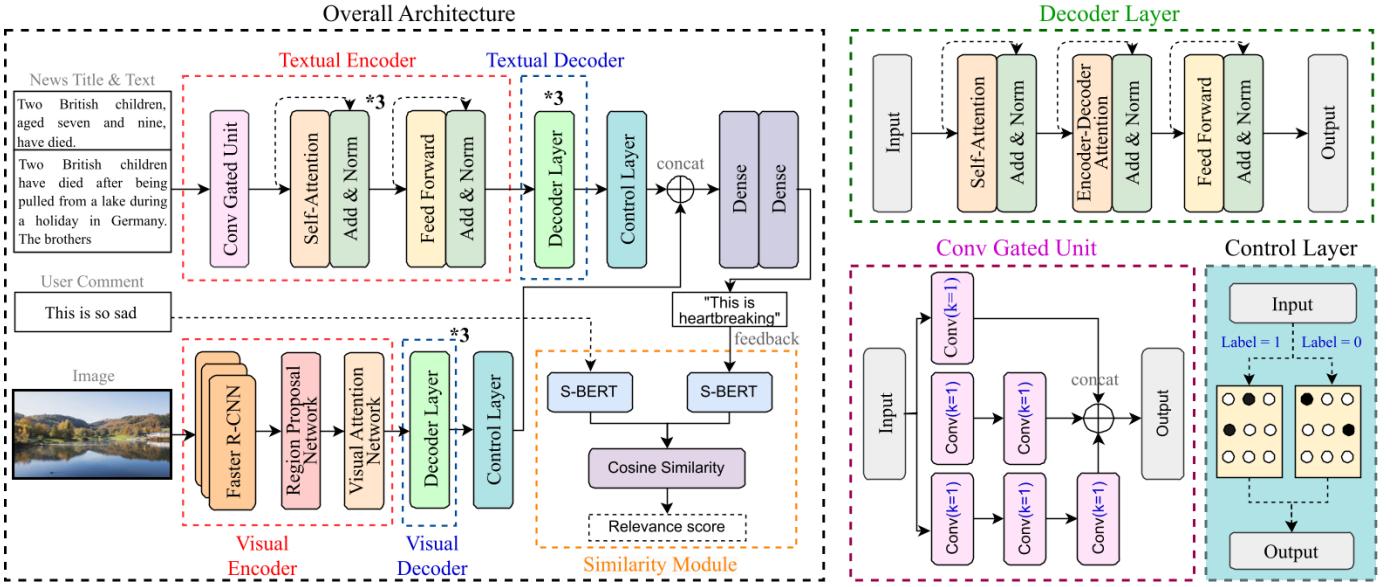


Fig. 1: The schematic architecture of the proposed system, encompassing encoder, decoder, and controllability blocks for textual and visual data. The decoder, convolution-gated unit, control layer, and similarity modules are delineated as distinct subblocks. Here, black circles denote off-neurons and white circles show on-neurons.

the total number of heads. These heads are then concatenated and multiplied by the weight matrix  $W$ , producing the intermediate output vector  $z'$ , as delineated in Eq. 4.

$$\begin{aligned} h_i &= \text{Attention}(QW^Q i, KW^K i, VW^V i) \\ z' &= \text{Concat}(h_1, h_2, \dots, h_{\text{head}})W^O \end{aligned} \quad (4)$$

where  $W^Q i$ ,  $W^K i$ ,  $W^V i$ , and  $W^O i$  are the respective projections of queries, keys, values, and output of corresponding heads. The final context vector  $z^*$  is then derived by passing this intermediate output through the feed-forward layer.

### 3.2.4 Decoder

The textual and visual decoders have an identical architecture comprising two main blocks: a self-attention block and an encoder-decoder attention block. These blocks are enhanced with positional encoding and normalization to improve efficiency and accuracy. They work separately. The textual decoder is fed with the textual context vector  $z_t$ , while the visual decoder receives the visual context vector  $z_i$ . Additionally, both decoders are provided with the ground-truth comment as input. The functioning of the attention layers is as follows:

- *Self-attention layer*: This layer utilizes future position masking to concentrate exclusively on prior positions in the output sequence. This layer's query, key, and value weight matrices are initially set to random values in the decoder and are progressively refined throughout the training process.
- *Encoder-decoder attention layer*: In this layer, queries are produced within the decoder. For the textual and visual context vectors, the context vectors  $z_t$  or  $z_i$  are used as keys and values in the matrices, respectively.

Late fusion has been applied using concatenation, preserving the distinct information of image features. A gated-convolution unit (GCU) has been introduced for textual feature encoding to minimize the repetition in the generated feedback.

### 3.2.5 Control Layer

The control layer is positioned following the decoder and just before the feedback generation phase. Its purpose is to introduce perturbations into the feedback synthesis model, thereby constraining the generated feedback to align with the desired sentiment. The control layer uses two masks, one for positive and the other for negative sentiments. These masks alter the input vector through element-wise multiplication, as described in Eq. 5. The role of the control layer is to fine-tune the sentiment in feedback so that it corresponds with the targeted sentiment tone. It accomplishes this by selectively activating or deactivating neurons to align the feedback with the chosen sentiment. This layer functions similarly to a modified dropout layer, designed to impact feedback sentiment through the regulation of neuron activity.

$$O = \begin{cases} mask_1 * I; & \text{Sentiment} = 0 \\ mask_2 * I; & \text{Sentiment} = 1 \end{cases} \quad (5)$$

where  $O$  and  $I$  denote the output and input vectors, respectively. Each mask in the setup blocks  $x\%$  of neurons, targeting different neuron sets. As a result,  $(100 - 2x)\%$  of neurons are trained for both sentiments, while  $x\%$  are specialized for a specific sentiment. This configuration directs the feedback towards the desired sentiment tone, with  $x\%$  set at 10% in this framework. The control mechanism trains  $(100 - 2x)\%$  of neurons on all samples for feature extraction, while  $x\%$  are dedicated to learning positive and negative sentiments in samples with corresponding sentiment tones. This method ensures the network's comprehensive understanding of both general and specific sentiment aspects in feedback. During inference, to generate sentiment-specific feedback, neurons trained for the contrasting sentiment are deactivated. For example, to produce positive feedback, neurons associated with negative sentiment are turned off, and vice versa. This approach is crucial for controlling the output sentiment independently of the input text's sentiment, focusing on steering the generated sentence's sentiment.

### 3.2.6 Similarity Module

The similarity module quantitatively assesses the semantic similarity between the feedbacks generated by the proposed system and human comments. It employs a pre-trained Sentence-BERT (SBERT) model [50] to generate sentence embeddings from the feedbacks and comments and use them for similarity comparisons. The individual comments and feedbacks are first transformed into vectors in an  $n$ -dimensional embedding space, where  $n$  denotes the size of the embeddings generated by the model. Each dimension within this space encapsulates a distinct linguistic feature or attribute, thereby comprehensively representing the sentence's semantic characteristics.

Following the generation of these embeddings, we compute the cosine similarity between the vector representations of the generated feedback and the human-provided comment. The cosine similarity score is a robust metric for this purpose, as it effectively captures the sentence vectors' orientation (and thus the semantic similarity) while remaining agnostic to their magnitude. It provides a relevance measure, reflecting the relevance and contextual similarity between the feedbacks and comments.

### 3.3 Interpretability

In this work, we have extended the K-Average Additive exPlanation (KAAP) interpretability technique introduced in our previous work [51]. It is based on Shapley Additive exPlanations (SHAP) that is an approximation of Shapley values [52]. As depicted in Fig. 2, it has been incorporated to assess the influence of each visual and textual feature on the sentiment portrayed by the generated feedback. We hypothesize that varying sentiments produced from identical inputs (text + images) should reflect in differential feature importance. It is expected that key features will differ for negative versus positive sentiments. When identical inputs are processed to portray varied sentiments, the model should adjust its focus across different image and text segments, thereby validating our controllability hypothesis.

**SHAP Values Computation:** The SHAP values for the features denote their contribution to the model's prediction. For a model  $f$ , the SHAP value for feature  $i$  is defined as per Eq. 6.

$$\mathcal{S}_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (6)$$

where  $F$  denotes the complete feature set,  $S$  a subset excluding  $i$ , and  $f(S)$  the model's prediction using features in  $S$ .

The computation of SHAP values requires exponential time theoretically which is approximated by dividing the input into  $k$  parts as illustrated in Eq. 7. For each modality, the input is repeatedly divided into  $k$  segments, determining each segment's impact on model predictions. A feature vector  $X$  with  $n$  features is segmented into  $k$  parts.

$$X = [X_1, X_2, \dots, X_k], \text{ where } X_i \subseteq X \text{ and } \bigcup_{i=1}^k X_i = X \quad (7)$$

For simplicity with  $k = 2$ , the fundamental computation of SHAP values is denoted in Eq. 8. It is extended for other values of  $k$ . The optimal values of  $k_{\text{img}}$  and  $k_{\text{txt}}$ , representing the number of segments to divide the image and text into, have been determined experimentally.

$$\mathcal{S}_{\{f_1\}} + \mathcal{S}_{\{f_2\}} = \mathcal{S}_{\{f_1, f_2\}} - \mathcal{S}_{\{\text{null}\}} \quad (8)$$

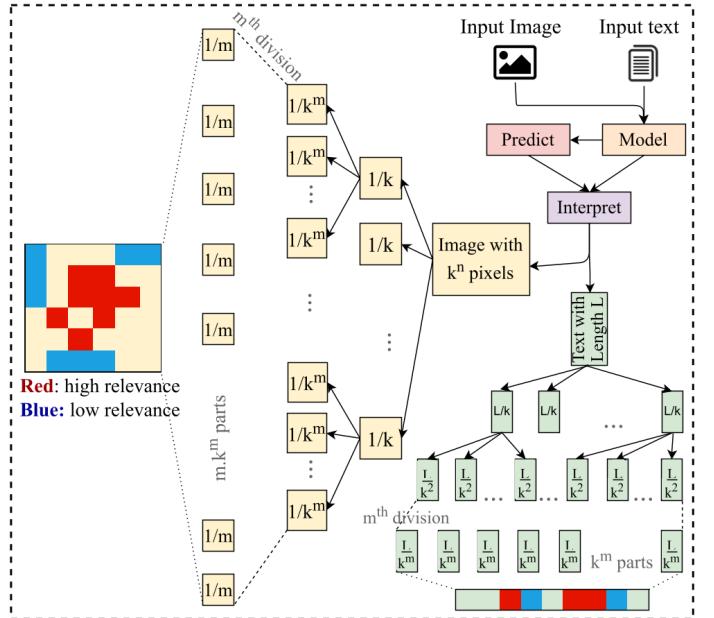


Fig. 2: Representation of the proposed interpretability technique. Here  $k_i$  and  $k_t$  denote the no. of partitions for image and text,  $w_i$  is image's width and  $L_t$  is text feature vector's length.

**K-Average Additive exPlanation (KAAP):** The KAAP value for feature  $i$  is calculated by averaging the SHAP values across the  $k$  divisions of the feature vector using Eq. 9.

$$\text{KAAP}_i = \frac{1}{k} \sum_{j=1}^k \mathcal{S}_i(X_j) \quad (9)$$

The KAAP values directly indicate the significant image features for predictions. For input image  $X^{\text{img}}$  of dimensions  $128 \times 128$ , the KAAP values for a given  $k$  are computed by segmenting the input along both axes. For text data  $X^{\text{txt}}$ , we derive the feature vector and divide it into  $k$  segments. Text division considers each word as a feature, acknowledging that sentiments are conveyed by words, not individual letters.

## 4 DATASET CONSTRUCTION

A large-scale dataset named controllable multimodal feedback synthesis (CMFeed) dataset<sup>1</sup>, has been constructed for controllable feedback synthesis. It contains 61734 comments from 3646 posts compiled by crawling news articles from Sky News, NYDaily, FoxNews, BBC News, and BBC NW through Facebook posts. The collection process utilized the NLTK<sup>2</sup> & newspaper3k<sup>3</sup> libraries.

The CMFeed dataset consists of multiple images per sample, corresponding news text, post likes and shares, and human comments. The comments for each post have been sorted based on Facebook's 'most-relevant' criterion, which prioritizes the comments with the highest likes and shares. The comments have been preprocessed in the following manner – the emoticons have been converted to words using the demoji library<sup>4</sup>; blank comments have been removed; contractions have been expanded; special

1. The CMFeed dataset and proposed system's code can be accessed at: [github.com/MIntelligence-Group/CMFeed](https://github.com/MIntelligence-Group/CMFeed).

2. <https://nltk.org/>

3. <https://newspaper.readthedocs.io/>

4. <https://pypi.org/project/demoji/>

TABLE 1: Representative samples from the CMFeed dataset. Here, ‘Post Likes’ and ‘Comment Likes’ show the number of likes for the post and comment, respectively. ‘Share’ denotes post shares and ‘Senti-class’ represents comment’s sentiment (1: positive, 0: negative).

Title	Text	Images	Post Likes	Shares	Comment	Comment Likes	Senti-Class
There are costs of managing beavers, but the benefits outweigh those costs.	Beaver dams in east Devon create area of wetland amid drought, The dams have created a wetland despite the dry weather. A network of dams built by beavers in Devon has helped to maintain an area of wetland despite a drought in the South West. There are a number of beaver families living on land....		2887	165	Of course the benefits outweigh the costs because beavers are ecosystem engineers!	47	1
A national emergency has been declared.	Pakistan floods: Monsoons bring misery to millions in PakistanBy Pumza Fihlani in Sukkur, Pakistan and Frances Mao in Singapore. Millions of people have been affected by floods in Pakistan, hundreds have been killed, and the government has declared a national emergency. On Friday, the National....		2005	126	Circumstances are really miserable after monster floodings, million of hectors crops have been destroyed and millions of people are homeless...	26	0
Celebrating his birthday, John Tinniswood said moderation in everything .....	Moderation is the key to life, GB's oldest man says on 110th birthday, Mr Tinniswood was joined by family and friends to celebrate his big day. Britain's oldest man has celebrated his 110th birthday by declaring "moderation in everything and all things" as the secret to a long and happy life. John Alfred.....		12000	407	Congratulations on a well lived life and 110	31	1
If it leaks then it's easy to clean.	Bin strikes: The people using baths and hiring skips to store rubbish, Helen Sikora has been keeping rubbish bags in her bath so she can easily clean up any leaks. Edinburgh residents have told how they have hired skips and even used bathtubs to store rubbish, as waste piles up during the city bin....		464	29	Dump it in the streets of they don't bother collecting it. We pay for a service.	29	0
...	...	...	...	...	...	...	...

and accented characters have been eliminated; punctuations and numbers have been purged to reduce noise; the stop-words have been removed, and the comments have been converted to lowercase. The preprocessing steps have been designed to strike a balance between reducing extraneous information (noise) and retaining crucial sentiment cues by the textual content. By converting emoticons to words, their sentiment value is preserved in a textual format that is more consistent for processing and analysis. This focus on textual content helps in creating a more consistent and analyzable dataset. While punctuations and numbers can carry sentiment information, their removal was aimed at reducing noise and enhancing the focus on textual content. This decision is based on the understanding that the core sentiment is primarily carried by words rather than punctuation or numbers.

To find the ground-truth sentiment labels for the comments, we first obtained sentiment scores using four pre-trained models: FLAIR Library [53], SentimentR [54], DistilBERT [55], and RoBERTa [56]. The outputs of these models included the sentiment value (-1 or 1) and the confidence for the sentiment. We first found the score as the multiplication of sentiment and confidence values. Then, we normalized the score to a value between 0 and 1 where 0 denoted negative and 1 denoted positive sentiment. Finally, the average of the normalized scores by these four models provided the final sentiment labels. A score between 0 and 0.5 denoted negative where 0 corresponded to most negative, whereas a score between 0.5 and 1 denoted positive sentiment class where 1 corresponded to most positive.

Various parameters of the CMFeed dataset have been described in Table 2 while sample data instances have been depicted in Fig.1. The CMFeed database differs from the existing datasets as

TABLE 2: CMFeed dataset’s parameters and respective values

Parameter	Value
No. of news posts	3646
No. of data samples	61734
Avg. no. of likes per post	65.1
Avg. no. of likes per comment	10.5
Avg. length of news text	655 words
Avg. no. of images per post	3.7
Percentage of positive comments	39.7

it consists of human-generated comments towards the image-text inputs along with the relevance scores for the comments. Multiple images and comments correspond to each news post, enabling the feedback synthesis model to learn the comments’ contextual diversity and relevance with the input. just

## 5 EXPERIMENTAL RESULTS

### 5.1 Training Protocols

The baselines and proposed system described in the following sections have been trained for 60 epochs on Nvidia V100 GPU. The experiments have been performed using 5-fold cross-validation and 80%-20% training-testing split using the following parameters.

- *General parameters* – Batch-size: 16, learning rate: 0.001, network optimizer: Adam, loss function: cross-entropy loss, activation function: ReLU.
- *Parameters for the transformer model* – Encoder embedding dimensions: 100, decoder embedding dimensions: 100, encoder hidden units dimensions: 128, decoder hidden units

- dimensions: 128, encoder dropout: 0.1, decoder dropout: 0.1, encoder no. of layers and attention heads: 3 and 8, decoder no. of layers & attention heads: 3 and 8, metric: accuracy.
- *Parameters for the Faster R-CNN model* – No. of epochs: 18, no. of proposals: 18, no. of anchor-box classes: 1601, network optimizer: adaDelta, metric: mAP (mean Average Precision).

## 5.2 Evaluation Metrics

Feedback synthesis is a one-to-many task, i.e., many feedbacks can be generated for one pair of image-text input. Hence, computing accuracy of generated feedbacks is not feasible. Instead, we evaluate the generated feedbacks against the ground-truth comments using the metrics to evaluate semantic relevance and their ranks. For semantic relevance evaluation, BLEU [57], CIDEr [58], ROUGE [59], SPICE [60], and METEOR [61] have been used. The BLEU is a precision-based metric, ROUGE is a recall-based metric, while METEOR combines precision and recall. On the other hand, CIDEr is a consensus-based evaluation metric, and SPICE is based on evaluating the sensitivity of n-grams. Higher values of these metrics denote more semantic similarity between the generated feedback and ground-truth comment. For rank based evaluation, ‘Recall@k’ [62] and ‘Mean Reciprocal Rank’ [63] have been used as per the following definitions. As we find top  $k$  results and evaluate them, these rank-based metrics are suitable for the evaluation.

- **Mean Reciprocal Rank:** For calculating Mean Reciprocal Rank (MRR), first, the similarity of the generated feedback is compared with all the ground-truth comments. If the most similar comment is ranked  $k$ , then the rank of the  $j^{th}$  feedback is given by Eq. 10.

$$rank_j = k \quad (10)$$

where  $k$  denotes the  $k^{th}$  comment when sorted by the relevance whereas  $rank_j$  is the Rank of the  $j^{th}$  feedback.

Finally, MRR is calculated as the average of the reciprocal ranks of all the generated feedback samples as per Eq. 11.

$$MRR = \left( \frac{1}{n} \right) \sum_{j=1}^n \frac{1}{rank_j} \quad (11)$$

where  $n$  is the number of generated feedback samples, while  $rank_j$  denotes the  $j^{th}$  feedback’s rank.

- **Recall@k:** Recall@k counts the number of data samples matching any top- $k$  relevant data samples. Adapting Recall@k for evaluating the generated feedback, the number of feedbacks similar to any of the top- $k$  comments sorted according to relevance is calculated. To find if the generated feedback is similar to any comment, the rank of that feedback as calculated in Eq. 11 is used. Finally, the Recall@k can be formulated according to Eq. 12. According to this, if the rank of the feedback is in top- $k$ , then a score of 1 is assigned to the feedback, else 0. The summation of all scores is done to calculate the Recall@k as shown in Eq. 12.

$$Recall@k_i = 1 \text{ if } rank_i \in [1, \dots, k] \\ Recall@k = \sum_{i=1}^n Recall@k_i \quad (12)$$

where  $Recall@k_i$  and  $rank_i$  is the Recall@k and rank of  $i^{th}$  feedback respectively whereas  $Recall@k$  denotes final score.

Furthermore, the sentiments of the generated feedbacks have been computed and sentiment classification accuracy has been analysed along with the ‘Control Accuracy’ which is the difference between the accuracies of controlled and uncontrolled feedbacks.

## 5.3 Baseline Models

The following baselines models have been constructed along with the proposed model. It is critical to highlight that the Controllability module, outlined in Section 3.2.5, is a consistent feature across all the subsequent architectural descriptions.

- **Baseline 1:** The first baseline utilizes Gated Recurrent Units (GRU) as textual encoder and VGG network as visual encoder. An early fusion method is applied to integrate visual and textual modalities.
- **Baseline 2:** Maintaining GRU for textual encoding and VGG for visual encoding, this baseline utilizes a late fusion approach for combining the visual and textual data.
- **Baseline 3:** This baseline implements a combination of a Transformer and a Gated Convolutional Unit (GCU) for textual encoding. It uses Faster RCNN with an additional visual attention mechanism for visual encoding. A late fusion strategy with averaging is used here.
- **Baseline 4:** The third baseline replaces the textual encoder with GPT-2 [64] while continuing to use Faster RCNN for visual data encoding with visual attention. It also combines the modalities using a late fusion with averaging. It has been empirically observed that GPT-2 based model generated good feedbacks only for textual input; however, it did not generate good feedbacks for multimodal input.
- **Proposed System:** The proposed system incorporates Transformer as the textual encoder and Faster RCNN as the decoder and it uses concatenation along with late fusion.

## 5.4 Results

The generated feedbacks have been evaluated quantitatively and qualitatively for a holistic assessment of their relevance with respect to ground-truth human comments.

**Semantic Relevance Evaluation:** The generated feedbacks’ semantic relevance with human comments has been evaluated. The feedbacks are generated to reflect the same sentiment class as reflected by the corresponding comments and then the feedbacks are evaluated using the BLEU, CIDEr, ROUGE, SPICE, and METEOR metrics. As depicted in Table 3, the proposed model has obtained the best values for these metrics in most cases.

TABLE 3: Quantitative evaluation of the feedbacks.

Model	BLEU	CIDEr	ROUGE	SPICE	METEOR
<b>Baseline 1</b>	0.1942	0.1342	0.2527	0.1028	0.0929
<b>Baseline 2</b>	0.2122	0.1635	0.2748	0.1654	0.1394
<b>Baseline 3</b>	0.2093	<b>0.1835</b>	0.2377	0.1555	0.1407
<b>Baseline 4</b>	0.1953	0.1798	0.2471	0.1478	0.1407
<b>Proposed</b>	<b>0.3020</b>	0.1817	<b>0.3378</b>	<b>0.1554</b>	<b>0.1412</b>

**Rank-based Evaluation:** The generated feedbacks are further evaluated using the MRR and Recall@k metrics. As observed in Table 4, 76.58% feedbacks are relevant to any top 10 comments. The MRR observed is 0.3789, denoting that most generated feedbacks are contentually similar to one of the top-3 comments.

The variations in sentiment classification accuracy and MRR varied differently for different models. For example, baseline 4 has lower MRR but high sentiment classification accuracy, whereas it

TABLE 4: Qualitative evaluation. Here, ‘MMR’ and ‘R@k’ denote ‘Mean Reciprocal Rank’ and ‘Recall@k’ where  $k \in \{1,3,5,10\}$ .

Model	MRR	R@1	R@3	R@5	R@10
<b>Baseline 1</b>	0.3435	17.30	39.67	60.67	67.75
<b>Baseline 2</b>	0.3305	17.69	36.99	<b>61.47</b>	74.29
<b>Baseline 3</b>	0.3214	16.08	37.53	59.32	69.29
<b>Baseline 4</b>	0.3182	16.98	37.26	56.11	71.29
<b>Proposed</b>	<b>0.3789</b>	<b>18.76</b>	<b>40.92</b>	60.13	<b>76.58</b>

is reverse for baseline 3. The proposed model provides the right trade-off with high values for both.

**Sentiment-Control:** Table 5 reports the Control Accuracies, which represent the difference in accuracies between controlled and uncontrolled feedbacks, for both the baselines and the proposed models. These models take the desired sentiment for the feedback to portray as one of the input parameters: 0 for negative and 1 for positive. In uncontrolled settings, this parameter is not used as the control layer is disabled. In contrast, controlled settings involve passing the ground-truth comment’s sentiment as the parameter.

TABLE 5: Synthesized feedbacks’ sentiment analysis. Here, ‘USentiAcc’ and ‘CSentiAcc’ denote the sentiment classification accuracies for uncontrolled and controlled feedbacks respectively.

Model	USentiAcc	CSentiAcc	Control Acc
<b>Baseline 1</b>	52.34	63.10	10.76
<b>Baseline 2</b>	54.72	67.06	12.34
<b>Baseline 3</b>	48.25	57.32	9.07
<b>Baseline 4</b>	52.48	71.57	<b>19.09</b>
<b>Proposed</b>	<b>58.41</b>	<b>77.23</b>	18.82

The sentiment class of the generated feedbacks is then computed using the FLAIR Library [53], SentimentR [54], DistilBERT [55], and RoBERTa [56], in a manner similar to that described in Section 4. The proposed model achieves a sentiment classification accuracy of 77.23% and a control accuracy of 18.82%. In order to calculate the sentiment accuracy, one negative and one positive feedback is generated by passing the respective parameter i.e. we passed parameter 0 for negative and 1 for positive feedback. Further, the sentiment of generated feedback is calculated and compares to the ground truth sentiment labels.

## 5.5 Ablation Studies

Following ablation studies have been conducted to evaluate the impact of various parameters on the proposed system’s performance.

### 5.5.1 Effect of Number of Control Layers and Value of Control-Parameter

The control layer has been used after the decoder and before the text generation phase to apply ‘control’ or constraints on the text generation. Here, it is crucial to decide a) the number of control layers and b) the suitable value for the control parameter. Regarding the number of control layers, we experimented with 1, 2, 3, and 4 control layers. The best performance has been observed using the 1 control layer, which decreased slightly for 2 control layers, decreased further for 3 layers, and decreased significantly for 4 layers. Further, regarding the control parameter, we experimented with its values of 5%, 10%, 15%, and 20%. The results show

that more control can be achieved with the increasing value of the control parameter; however, fewer neurons will get trained for the entire training data, causing a degradation in the result quality. As depicted in Table 6, a control value of 10% results in better performance than the other values. Hence, 1 control layer with control value of 10% have been used in the final implementation.

### 5.5.2 Effect of Beam-size

The beam-size is a search parameter that refers to the number of options the model keeps at each step of the prediction, controlling the breadth of the search for the best output sequence. It keeps only the top  $k$  predictions, where  $k$  is the beam size. A larger beam size allows the model to explore more possibilities, potentially improving output quality; however, it increases computational requirements and may also degrade the output because of repetitive text generation. We experimented with beam-size values of 2, 5, 10, 15, and 20. The corresponding sentiment classification accuracies and MRR values have been summarized in Table 6. As the beam-size value of 5 provides the best performance and computational complexity trade-off, it has been used in the final implementation.

### 5.5.3 Effect of Division Factor for KAAP technique

The suitable values of the division factors  $k_{img}$  and  $k_{txt}$  used in Section 3.3 have been decided experimentally using the dice coefficient [65]. It measures the similarity of two data samples; the value of 1 denotes that the two compared data samples are completely similar, whereas a value of 0 denotes their complete dis-similarity. For each modality, we computed the KAAP values at  $k \in \{2, 3, \dots, 30\}$  and analyzed the dice coefficients for two adjacent  $k$  values. For image & text, the dice coefficient values converge to 1 at  $k$  values of 5 and 20 respectively. Hence, the same have been used by the proposed system.

## 5.6 Visualization and Interpretability

Fig. 3 shows sample results, highlighting the features being focused on during uncontrolled and controlled feedbacks’ generation. In image plots, red and blue represent the most and least contributing pixels, respectively whereas for text plots, yellow and blue indicates the most and least important textual features.

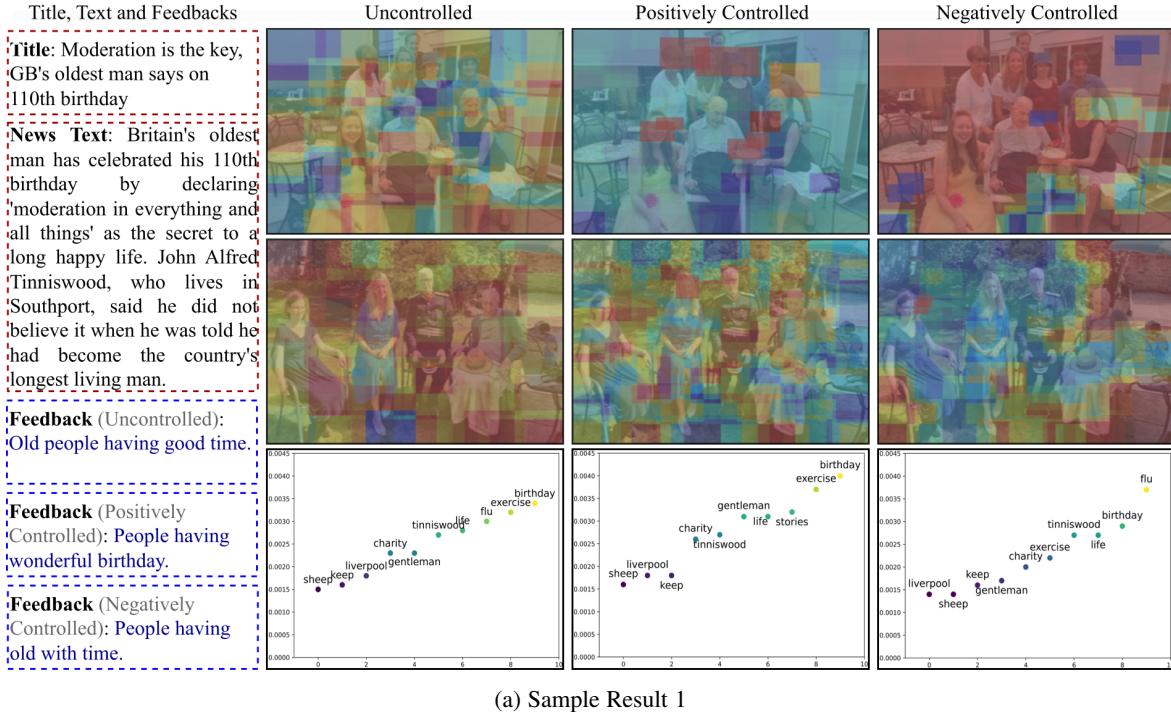
In Fig. 3a, positive sentiments are indicated by smiling faces and a family setting, while negative sentiments are associated with the depiction of aging, particularly in the older face. The expression of the girl on the left, a blend of smiling and discomfort, captures attention in both positive and negative contexts. The middle girl’s face, predominantly smiling, is highlighted in red for positive and blue for negative sentiments. Fig. 3b shows that dark areas contribute to negative sentiment, whereas faces are linked to positive sentiment. In negatively controlled settings, the crowd is focused; in positively controlled settings, the focus shifts to individual people. In Fig. 3c, positive sentiments downplay the importance of the gun, concentrating more on the number plate. In the uncontrolled setting, the focus is primarily on the words. For Fig. 3d, the facial features are highlighted red for positively controlled and blue for negatively controlled settings for the first image. The second image associates positive sentiment with light and text and negative sentiment with darkness.

## 5.7 Human Evaluation

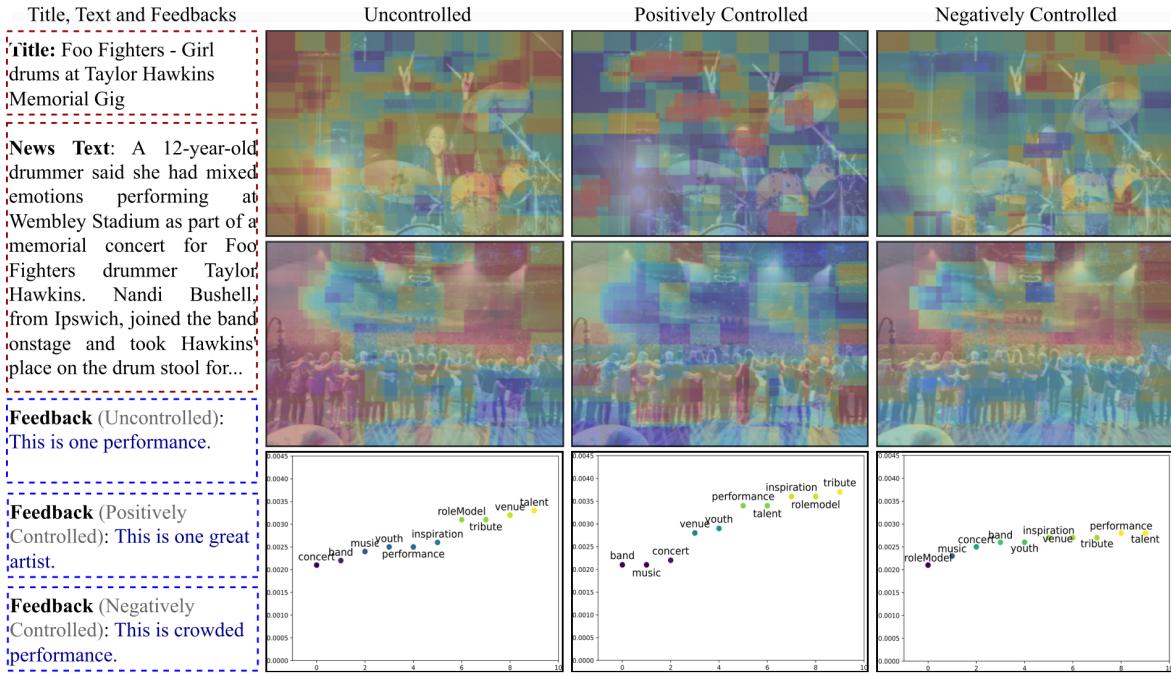
The sentiments of the generated feedbacks have been evaluated by 50 evaluators, comprising 25 males and 25 females, with an

TABLE 6: Ablation studies on control parameter ( $x$ ) and beam-size. The entries show sentiment classification accuracy / MRR.

$x$ / Beam-size	2	5	10	15	20
<b>5</b>	66.92% / 0.3505	76.89% / 0.3605	52.83% / 0.3641	54.51% / 0.3214	50.40% / 0.3491
<b>10</b>	71.42% / 0.3483	<b>77.23% / 0.3789</b>	64.81% / 0.3390	50.48% / 0.3393	47.36% / 0.3503
<b>15</b>	52.82% / 0.3429	69.71% / 0.3548	54.99% / 0.3312	60.40% / 0.3523	45.24% / 0.3449
<b>20</b>	64.57% / 0.3354	75.12% / 0.3389	57.76% / 0.3355	49.63% / 0.3409	41.46% / 0.3295



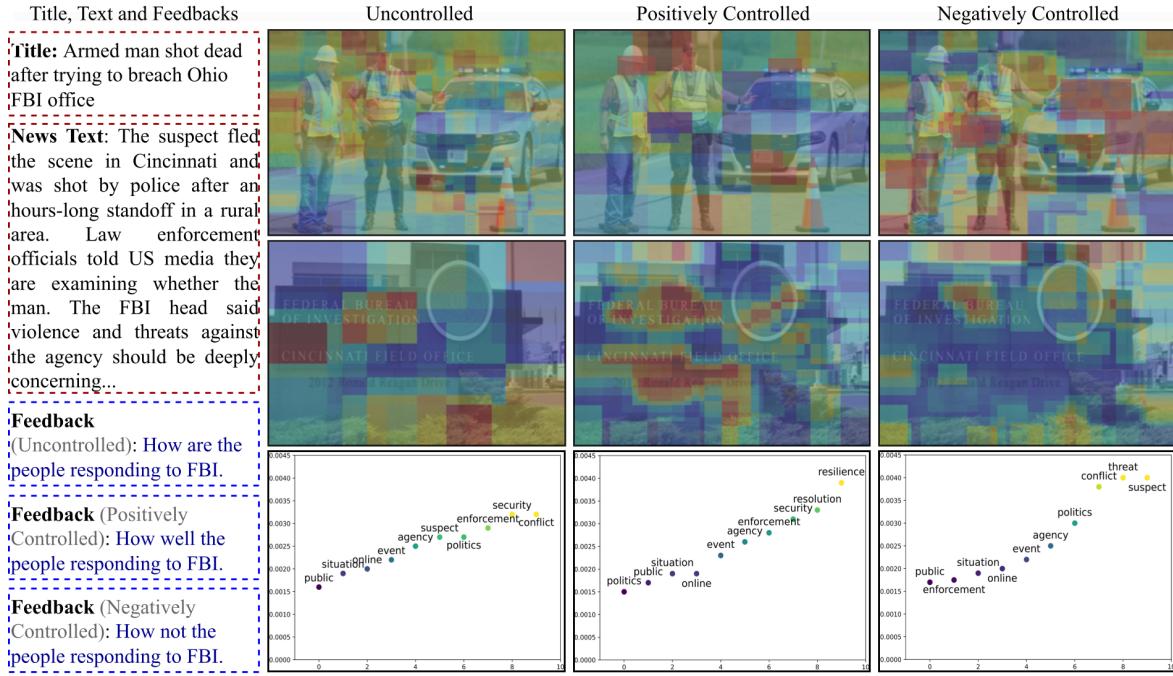
(a) Sample Result 1



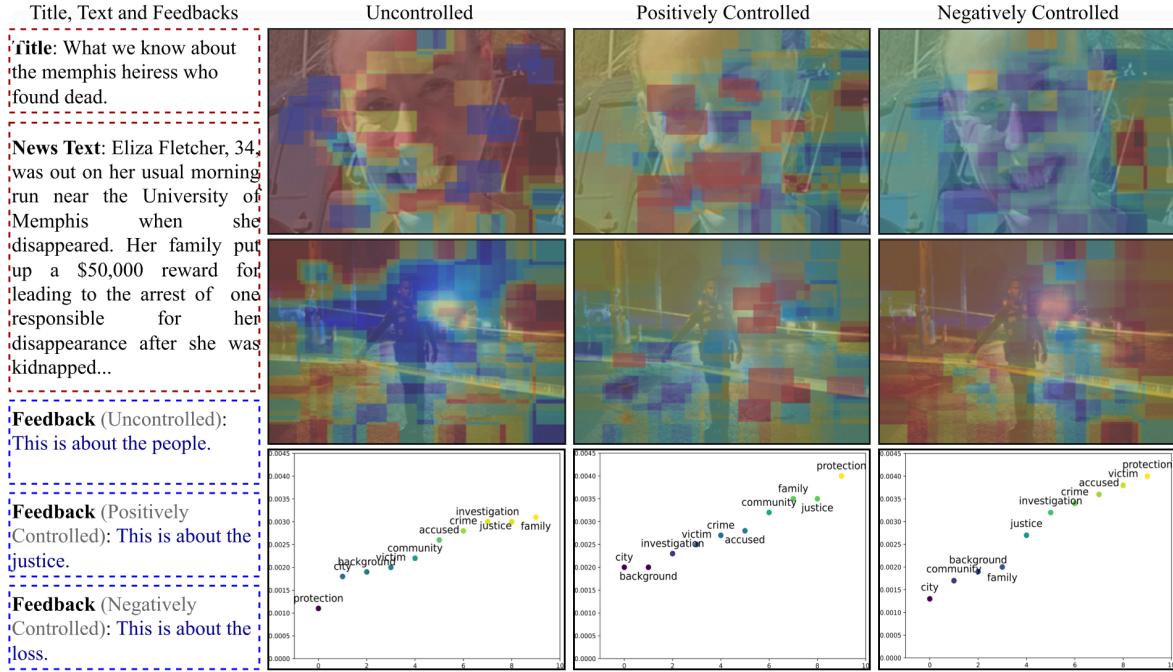
(b) Sample Result 2, continued on next page.

average age of  $30 \pm 2.73$  years. They assessed the controlled and uncontrolled feedbacks for their valence and relevance with the inputs. The samples have been picked randomly, and the averages of the evaluators' scores have been reported. These scores for

the relevance ratings for three types of feedback—uncontrolled, positive, and negative—against images and texts have been described in Table 7. On average, 72.68% and 78.14% evaluators reported that the sentiments of positively and negatively controlled



(c) Sample Result 3



(d) Sample Result 4

Fig. 3: Sample results along with interpretability plots. They depict the feedback generated by the proposed system using the news headline, text, and images (two out of multiple images shown) under the given sentiment-controllability constraint.

feedbacks are more positive and more negative respectively than uncontrolled feedbacks'. The higher relevance scores for controlled feedbacks ( $F_{PosCtrl}$  and  $F_{NegCtrl}$ ) compared to uncontrolled ones ( $F_{UnCtrl}$ ) confirm the control layer's influences in feedback's desired sentiment alignment.

## 6 DISCUSSION AND CONCLUSIONS

The proposed system, trained on human comments and image-text inputs under sentiment constraints, generates human-like

feedback with appropriate sentiments, as evidenced by evaluation metrics in Tables 3 and 4. It effectively manages sentiment in feedback, excluding negative content such as 'hatred,' and produces sentiment-specific responses, simulating human sentiment in context. The control layer allows for feedback generation with desired sentiments, utilizing different non-keywords and varying keywords, especially with higher control parameter values. The challenge of determining comment relevance in the CMFeed dataset was addressed by adopting Facebook's relevance criteria.

TABLE 7: Human evaluation of generated feedbacks where  $F_{UnCtrl}$ ,  $F_{PosCtrl}$  and  $F_{NegCtrl}$  denote uncontrolled, positively controlled and negatively controlled feedbacks.  $Rel_{img}$ ,  $Rel_{text}$ ,  $Rel_{Comment}$  and  $Rel_{F_{UnCtrl}}$  denote the ‘relevance with’ input images, text, comments and the uncontrolled feedback, respectively.

	$Rel_{img}$	$Rel_{text}$	Comment	$Rel_{F_{UnCtrl}}$
<b>Comment</b>	70.85%	72.93%	100.00%	78.27%
<b>F<sub>UnCtrl</sub></b>	67.27%	69.58%	78.27%	100.00%
<b>F<sub>PosCtrl</sub></b>	69.47%	71.07%	79.93%	81.96%
<b>F<sub>NegCtrl</sub></b>	71.23%	72.13%	80.17%	83.24%

Notably, we pioneered the task of feedback synthesis in our previous work [1], and here we have built upon it to generate controllable and interpretable feedbacks. The current work achieves comparable but slightly lower ‘R@k’ values, justifiable by the larger number of comments per sample (10.5 compared to 8.5). However, it achieves a higher MRR value (0.3789 versus 0.3042), suggesting the effectiveness of the similarity module implemented in this work, especially given the higher comment count per sample.

The model-level and case-level interpretability analysis has been incorporated. Model-level interpretability is achieved by introducing perturbations to the feedback synthesis model via the control layer. The impact of these perturbations on the output feedback is detailed in Table 5 in terms of sentiment classification accuracies for uncontrolled and controlled (perturbed) scenarios. Additionally, Fig. 3 exemplifies case-level interpretability, illustrating how the model’s output varies in different situations.

The sentiment analysis is based on the outputs from four pre-trained models: FLAIR Library, SentimentR, DistilBERT, and RoBERTa. The choice of a binary classification system is influenced by the fact that some of these models inherently provide binary outputs (positive and negative, without neutral). The primary focus of this task is to demonstrate the controllability of the sentiments portrayed by the generated feedback towards the multimodal input, and our work is the first to do that. In the future, we plan to explore discrete emotion classes, integrate additional modalities, such as audio or physiological signals, to enhance the richness and accuracy of sentiment-controlled feedback synthesis.

## ACKNOWLEDGEMENTS

The authors acknowledge the CSC-IT Center for Science, Finland, for providing computational resources.

## REFERENCES

- [1] P. Kumar *et al.*, “Affective Feedback Synthesis Towards Multimodal Text and Image Data,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–23, 2023.
- [2] M. R. Makiuchi, K. Uto *et al.*, “Multimodal Emotion Recognition with High-level Speech and Text Features,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.
- [3] F. R. Gallo, G. I. Simari *et al.*, “Predicting User Reactions to Twitter Feed Content based on Personality Type and Social Cues,” *Future Generation Computer Systems*, vol. 110, pp. 918–930, 2020.
- [4] M. Muszynski *et al.*, “Recognizing Induced Emotions of Movie Audiences from Multimodal Information,” *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 36–52, 2019.
- [5] P. Blkstein and M. Worsley, “Multimodal Learning analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks,” *Journal of Learning Analytics*, vol. 3, no. 2, pp. 220–238, 2016.
- [6] A. Vaswani, N. Shazeer *et al.*, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [7] S. Ren, K. He *et al.*, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 91–99, 2015.
- [8] S. Gao, X. Chen *et al.*, “From Standard Summarization to New Tasks and Beyond: Summarization With Manifold Information,” in *The 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.
- [9] X. Gu, K. Cho *et al.*, “DialogWAE: Multimodal Response Generation With Conditional Wasserstein Auto Encoder,” in *The International Conference on Learning Representations (ICLR)*, 2019.
- [10] S. Antol, A. Agrawal *et al.*, “VQA: Visual Question Answering,” in *The 19th IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.
- [11] A. Das, S. Kottur *et al.*, “Learning Cooperative Visual Dialog Agents With Deep Reinforcement Learning,” in *The 21th IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2951–2960.
- [12] H. Zhou, M. Huang *et al.*, “Emotional Chatting Machine: Emotional Conversation Generation With Internal And External Memory,” in *AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [13] Y. Wu, F. Wei *et al.*, “Response Generation by Context Aware Prototype Editing,” in *The 33rd AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 7281–7288.
- [14] A. Radford, K. Narasimhan *et al.*, “Improving Language Understanding by Generative Pre-training,” *OpenAI*, 2018.
- [15] Y. Zhu, W. Zhao *et al.*, “Topic-Aware Video Summarization Using Multimodal Transformer,” *Pattern Recognition*, vol. 140, p. 109578, 2023.
- [16] J. Chen and H. Zhuge, “Abstractive Text-Image Summarization Using Multimodal Attention Hierarchical RNN,” in *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 4046–4056.
- [17] J. Zhu, H. Li *et al.*, “MSMO: Multimodal Summarization With Multimodal Output,” in *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 4154–4164.
- [18] R. Zellers *et al.*, “VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles,” in *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9360–9369.
- [19] X. Shang, Z. Yuan *et al.*, “Multimodal Video Summarization Via Time-Aware Transformers,” in *29th ACM International Conference on Multimedia*, 2021, pp. 1756–1765.
- [20] B. Zhao, M. Gong, and X. Li, “Audio-Visual Video Summarization,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [21] J. Xie, X. Chen *et al.*, “Multimodal-Based And Aesthetic-Guided Narrative Video Summarization,” *IEEE Transactions on Multimedia*, 2022.
- [22] M. Page Fortin and B. Chaib-draa, “Multimodal Multitask Emotion Recognition Using Images, texts and tags,” in *The ACM International Conference on Multimedia Retrieval (ICLR)*, 2019, pp. 3–10.
- [23] J. Zhu, Y. Zhou *et al.*, “Multimodal Summarization With Guidance of Multimodal Reference,” in *The 34th AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 05, 2020, pp. 9749–9756.
- [24] F. Chen, J. Xie *et al.*, “Graph Convolutional Net For Difficulty-Controllable Visual Ques. Generation,” *World Wide Web*, pp. 1–23, 2023.
- [25] P. Wang, Q. Wu *et al.*, “FVQA: Fact-Based Visual Question Answering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2413–2427, 2017.
- [26] P. Cascante-Bonilla, H. Wu *et al.*, “SimVQA: Exploring Simulated Environments for Visual Question Answering,” in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5056–5066.
- [27] Z.-X. Jin, H. Wu *et al.*, “RUArt: A Novel Text-Centered Text-Based Visual Question Answering,” *IEEE Transactions on Multimedia*, 2021.
- [28] Z. Guo, J. Zhao *et al.*, “A Universal Quaternion Hypergraph For Multimodal VQA,” *IEEE Transactions on Multimedia*, 2021.
- [29] J. Wu, T. Mu *et al.*, “Memory-Aware Attentive Control For Community Question Answering With Knowledge-Based Dual Refinement,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.
- [30] J. Lehmann *et al.*, “Language Models As Controlled Natural Language Semantic Parsers For Knowledge Graph Question Answering,” in *European Conference on Artificial Intelligence (ECAI)*, vol. 372. IOS Press, 2023, pp. 1348–1356.
- [31] G.-C. Kang, J. Lim *et al.*, “Dual Attention Networks for Visual Reference Resolution in Visual Dialog,” in *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 2024–2033.
- [32] X. Jiang *et al.*, “DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue,” in *The 34th AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 11125–11132.
- [33] X. Xu, O. Dušek *et al.*, “Better Conversations By Modeling, Filtering, And Optimizing For Coherence And Diversity,” *arXiv preprint arXiv:1809.06873*, 2018, Accessed 2023-01-31.

- [34] T. Zhao *et al.*, “Learning Discourse-Level Diversity For Neural Dialog Models Using Conditional Variational Autoencoders,” in *The 55th Annual Meeting of Association for Comp. Linguistics (ACL)*, 2017, pp. 654–664.
- [35] F. Chen, X. Chen *et al.*, “Improving Cross-Modal Understanding In Visual Dialog Via Contrastive Learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7937–7941.
- [36] Z. Wang, J. Wang *et al.*, “Unified Multimodal Model With Unlikelihood Training For Visual Dialog,” in *The 30th ACM International Conference on Multimedia*, 2022, pp. 4625–4634.
- [37] G.-C. Kang, S. Kim *et al.*, “The Dialog Must Go On: Improving Visual Dialog via Generative Self-Training,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6746–6756.
- [38] A.-A. Liu, G. Zhang *et al.*, “Closed-Loop Reasoning With Graph-Aware Dense Interaction For Visual Dialog,” *Multimedia Systems*, vol. 28, no. 5, pp. 1823–1832, 2022.
- [39] A.-A. Liu, C. Huang *et al.*, “Counterfactual Visual Dialog: Robust Commonsense Knowledge Learning From Unbiased Training,” *IEEE Transactions on Multimedia*, 2023.
- [40] W. Shi and Z. Yu, “Sentiment Adaptive End-to-End Dialog Systems,” in *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 1509–1519.
- [41] X. Kong, B. Li *et al.*, “An Adversarial Approach To Sentiment-Controlled Neural Dialogue Generation,” *arXiv preprint arXiv:1901.07129*, 2019, Accessed 2023-01-31.
- [42] M. Firdaus, H. Chauhan *et al.*, “EmoSen: Generating Sentiment And Emotion Controlled Responses In A Multimodal Dialogue System,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1555–1566, 2020.
- [43] M. Firdaus, U. Jain *et al.*, “SEPRG: Sentiment Aware Emotion Controlled Personalized Response Generation,” in *14th International Conference on Natural Language Generation*, 2021, pp. 353–363.
- [44] J. Hu, Y. Huang *et al.*, “The Acoustically Emotion-Aware Conversational Agent With Speech Emotion And Empathetic Responses,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 17–30, 2022.
- [45] T. Saha, S. Saha *et al.*, “Towards Sentiment-Aware Multi-Modal Dialogue Policy Learning,” *Cognitive Computation*, pp. 1–15, 2022.
- [46] K. Wang and X. Wan, “SentiGAN: Generating Sentimental Texts Via Mixture Adversarial Networks,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 4446–4452.
- [47] C. Huang, O. R. Zaiane *et al.*, “Automatic Dialogue Generation With Expressed Emotions,” in *The Conference of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018, pp. 49–54.
- [48] P. Zhong *et al.*, “An Affect-Rich Neural Conversational Model With Biased Attention And Weighted Cross-Entropy Loss,” in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7492–7500.
- [49] A. Rosebrock, “Intersection Over Union (IoU) for Object Detection,” *PyImageSearch.com*, 2016, Accessed 2023-01-31.
- [50] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.
- [51] P. Kumar, S. Malik *et al.*, “Hybrid Fusion Based Interpretable Multimodal Emotion Recognition With Insufficient Labelled Data,” *arXiv preprint arXiv:2208.11450*, 2022, Accessed 2023-01-31.
- [52] L. Shapley, “A Value for n-Person Games, Contributions to the Theory of Games II,” 1953.
- [53] A. Akbik, T. Bergmann *et al.*, “FLAIR: An Easy-to-Use Framework For State-Of-The-Art NLP,” in *2019 Conf. of North American Chapter of Association for Comp. linguistics (NAACL)*, 2019, pp. 54–59.
- [54] T. Rinker. (2017) Sentimentr Package for R Language. <https://github.com/trinker/sentimentr>. Accessed 2023-01-31.
- [55] V. Sanh, L. Debut *et al.*, “DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” *arXiv preprint arXiv:1910.01108*, 2019, Accessed 2023-01-31.
- [56] Y. Liu, M. Ott *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv preprint arXiv:1907.11692*, 2019, Accessed 2023-01-31.
- [57] K. Papineni, S. Roukos *et al.*, “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *The 40th Annual Meeting on Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
- [58] R. Vedantam, C. Lawrence Zitnick *et al.*, “CIDEr: Consensus-based Image Description Evaluation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
- [59] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*, 2004, pp. 74–81.
- [60] P. Anderson, B. Fernando *et al.*, “SPICE: Semantic Propositional Image Caption Evaluation,” in *The European Conference on Computer Vision (ECCV)*, 2016, pp. 382–398.
- [61] A. Lavie and M. J. Denkowski, “The METEOR Metric for Automatic Evaluation of Machine Translation,” *Springer Machine Translation Journal*, vol. 23, no. 2-3, pp. 105–115, 2009.
- [62] P. Runeson, M. Alexandersson, and O. Nyholm, “Detection of Duplicate Defect Reports using Natural Language Processing,” in *IEEE International Conference on Software Engineering*, 2007, pp. 499–510.
- [63] N. Craswell, “Mean Reciprocal Rank,” *Encyclopedia of Database Systems*, vol. 1703, 2009.
- [64] OpenAI. (2019) GPT2. [huggingface.co/docs/transformers/model\\_doc/gpt2](https://huggingface.co/docs/transformers/model_doc/gpt2). Accessed 2023-01-31.
- [65] R. Deng, C. Shen *et al.*, “Learning To Predict Crisp Boundaries,” in *The European Conference on Computer Vision (ECCV)*, 2018, pp. 562–578.



**Puneet Kumar** (Member, IEEE) received B.E. and M.E. degrees in Computer Science in 2014 and 2018, respectively, followed by a Ph.D. from the Indian Institute of Technology Roorkee, India in 2022. He has worked at Oracle Corporation, Samsung R&D, and PaiByTwo Pvt. Ltd., and is now a Postdoctoral Researcher at the University of Oulu, Finland. His research focuses on Affective Computing, Multimodal and Interpretable AI, Mental Health and Cognitive Neuroscience. He has published in top journals and conferences

and received various awards, including the institute medal in M.E., the best thesis, and several best paper awards. For more information, visit his personal website at [www.puneetkumar.com](http://www.puneetkumar.com).



**Sarthak Malik** received his B.Tech degree in Electrical Engineering from the Indian Institute of Technology Roorkee, India. He is currently a Data Scientist at MasterCard. He is an avid programmer with active participation in coding competitions and development projects. His research focuses on Affective Computing, Machine Learning, Computer Vision, Natural Language Processing and Interpretable AI with publications in reputed journals and conferences. He has achieved notable recognitions including a rank

in top 10 in Geoffrey Hinton Fellowship Hackathon. For more details, visit his webpage at [www.linkedin.com/in/sarthak-malik](http://www.linkedin.com/in/sarthak-malik).



**Balasubramanian Raman** (Senior Member, IEEE) received his B.Sc. and M.Sc. degrees from the University of Madras in 1994 and 1996, respectively, and his Ph.D. from the Indian Institute of Technology Madras, India in 2001. He is a Chair Professor in the Department of Computer Science and Engineering and a Joint Faculty in the Mehta Family School of Data Science and Artificial Intelligence at the Indian Institute of Technology Roorkee, India. He has published over 200 research papers in reputed journals and

conferences. His research interests include Machine Learning, Image Processing, Computer Vision, Activity & Pattern Recognition, Medical Imaging, Privacy-Preserving Computing, and Data Security. For more information, visit his webpage at <http://faculty.iitr.ac.in/cs/bala>.



**Xiaobai Li** (Senior Member, IEEE) received her B.Sc. and M.Sc. degrees in 2004 and 2007, respectively, and her Ph.D. from the University of Oulu, Finland, in 2017. She is currently a ZJU100 professor with the School of Cyber Science and Technology, Zhejiang University, and she is also an Adjunct Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. Her research focuses on Affective Computing, Facial Expression Recognition, Micro-Expression Analysis, Remote Physiological Signal Measure-

ment, and Healthcare. She co-chaired several international workshops in CVPR, ICCV, FG, and ACM Multimedia and is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology, Frontiers in Psychology, and Image and Vision Computing. For more information, visit her webpage at [www.oulu.fi/en/researchers/xiaobai-li](http://www.oulu.fi/en/researchers/xiaobai-li).