Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 9267487, 15 pages
https://doi.org/10.1155/2023/9267487

Research Article

Beyond Words: An Intelligent Human-Machine Dialogue System with Multimodal Generation and Emotional Comprehension

Yaru Zhao,[1] **Bo Cheng**,[1] **Yakun Huang**,[1] **and Zhiguo Wan**[2]

[1] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *Zhejiang Lab, Hangzhou 311121, China*

Correspondence should be addressed to Yaru Zhao; zhaoyaru@bupt.edu.cn

Received 18 April 2023; Revised 28 August 2023; Accepted 27 November 2023; Published 23 December 2023

Academic Editor: Alexander Hošovský

Intelligent service robots have become an indispensable aspect of modern-day society, playing a crucial role in various domains ranging from healthcare to hospitality. Among these robotic systems, human-machine dialogue systems are particularly noteworthy as they deliver both auditory and visual services to users, effectively bridging the communication gap between humans and machines. Despite their utility, the majority of existing approaches to these systems primarily concentrate on augmenting the logical coherence of the system's responses, inadvertently neglecting the significance of user emotions in shaping a comprehensive communication experience. To tackle this shortcoming, we propose the development of an innovative human-machine dialogue system that is both intelligent and emotionally sensitive, employing multimodal generation techniques. This system is architecturally comprised of three components: (1) data collection and processing, responsible for gathering and preparing relevant information, (2) a dialogue engine, which generates contextually appropriate responses, and (3) an interaction module, responsible for facilitating the communication interface between users and the system. To validate our proposed approach, we have constructed a prototype system and conducted an evaluation of the performance of the core dialogue engine by utilizing an open dataset. The results of our study indicate that our system demonstrates a remarkable level of multimodal generation response, ultimately offering a more human-like dialogue experience.

1. Introduction

A human-like dialogue system, characterized by its capacity for autonomous interaction and the ability to perceive and express emotions, has become increasingly relevant in today's technologically driven world [1]. Despite significant advancements in digital service robots, a majority of these systems still lack the required intelligence and emotional generation capabilities essential for enabling comprehensive multimodal human-machine interactions. Traditional dialogue systems, which are designed to generate responses to input text, primarily rely on advanced natural language processing techniques. Two of the most prominent techniques include sequence-to-sequence (Seq2Seq) models [2] and the innovative transformer architecture [3]. Both approaches have proven to be effective in generating responses; however, certain limitations exist when using these techniques in isolation. As illustrated in Figure 1, when dialogue systems depend exclusively on textual input, they often generate responses that are not only repetitive but also lacking in depth and engagement.

To address this shortcoming, the current research proposes the integration of both textual and emotional information in the dialogue system. In the realm of academia, researchers have extensively investigated dialogue models, such as those presented in Shuster et al. [4, 5], and have proposed emotion-enhanced models, as discussed in Wei et al. [6] and Li et al. [7]. Specifically, to address the limitations of single text generation models, multimodal dialogue models capable of processing both textual and video

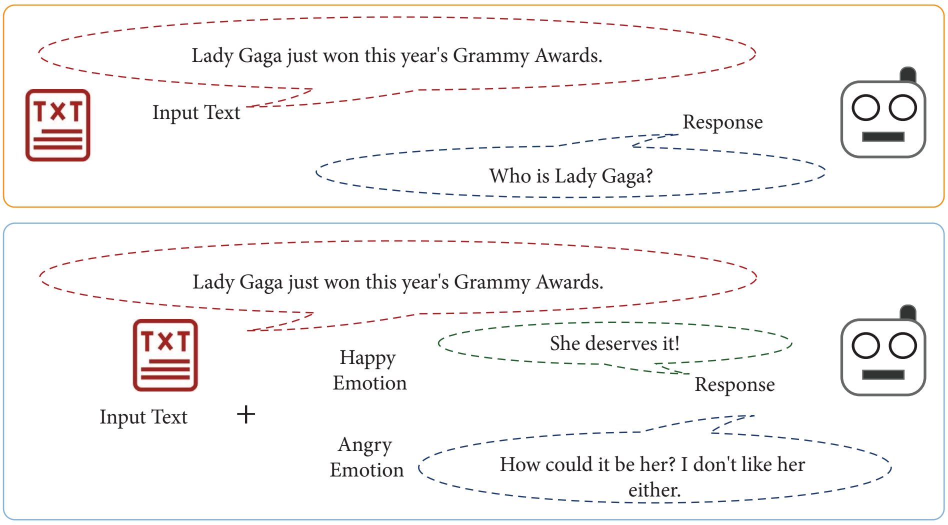
Lady Gaga just won this year's Grammy Awards.
TXT
Input Text
Response
Who is Lady Gaga?
Lady Gaga just won this year's Grammy Awards.
TXT
Input Text
+
Happy Emotion
She deserves it!
Response
Angry Emotion
How could it be her? I don't like her either.

Figure 1: Illustration of the text-based and multimodal dialogues, which is an original figure created by us.

information have been proposed, including the works of Fung et al. [8], Huber et al. [9], and Tian et al. [10]. More importantly, Shen et al. [11] designed ViDA-MAN, a digital human agent for multimodal interaction, which provides real-time audiovisual responses to users through voice queries. When investigating the development of the industry, intelligent dialogue service robots have been implemented across a diverse range of sectors, including industrial, domestic, medical, military, educational, and entertainment applications. Notable examples include Microsoft's Ice system, which demonstrates an understanding of emotional contexts to a certain extent, and Turing Robotics' AI robot operating system, which offers multimodal interaction modes. Additionally, Baidu has developed a sophisticated multimodal intelligence platform that integrates voice recognition, semantic understanding, face recognition, and gesture recognition, facilitating seamless human-computer interaction.

Despite these significant strides, there remains a pressing need for the development of dialogue systems that can effectively perform semantic fusion of multimodal inputs, ensuring a more cohesive and intuitive human-machine interaction experience. In this paper, we introduce an advanced human-machine dialogue system that exhibits emotional intelligence and employs deeply multimodal generation techniques to create a more human-like interactive experience. Two paramount challenges associated with the development of our dialogue engine arise:

(i) *Designing a Multimodal Dialogue System Integrating User Emotions*. Conventional dialogue systems generally accept text or voice inputs, generate responses using rule-based or generative models that rely on a knowledge base, and offer feedback in the form of text or voice. To integrate the user's emotions into such a system, it is necessary to modify the input, preprocessing, dialogue generation, and user interaction response mechanisms. Thus, the development and implementation of a multimodal dialogue system that effectively incorporates user emotions is a significant and challenging endeavor. In particular, the accurate detection and computation of user emotions, as well as providing more engaging feedback to the user through generated responses, are of paramount importance.

(ii) *Extracting Emotional and Semantic Features of Multimodal Inputs and Fusing Them for Human-Like Responses*. Different modal inputs provide varying semantic information. For example, a user's facial expressions may not change significantly, but their voice may convey dissatisfaction. Consequently, an emotion-aware multimodal dialogue generation cannot rely solely on a single visual emotion or a simplistic superposition approach. The extraction of emotional expressions from diverse modal inputs and their effective incorporation into the multimodal dialogue generation model to produce responses that reflect the user's emotions are essential in achieving a more human-like digital human interaction.

To address the first challenge, we developed a dialogue system that incorporates multimodal inputs, consisting of three primary components: data collection and processing, a dialogue engine, and interaction modules. The first component, the data collection and processing module, addresses the challenge of gathering and preprocessing information from diverse modalities, such as text, images, and audio. This module is also responsible for extracting the corresponding emotional features from the data, allowing for a more empathetic and context-aware dialogue system. The second component, the dialogue engine module, serves as the central processing unit of the dialogue system. It generates dialogues by integrating the original multimodal inputs and emotional features. This design ensures that the generated dialogues are not only contextually accurate but also emotionally coherent, thus providing a more realistic and engaging conversational experience. Lastly, the interaction module is designed to deliver the generated responses to the user in a captivating and immersive manner. It includes rendering and interaction steps that enable the dialogue system to respond to the user through a virtual digital human. This innovative approach allows the system to interact with the user as a realistic virtual character, enhancing the overall user experience. To facilitate seamless communication, the generated responses are fed back to the user via video streaming, incorporating elements from the digital humanities to create a more relatable and interactive virtual environment.

To address the second challenge, we introduce a novel multimodal dialogue generation model, which builds upon the transformer architecture by incorporating an encoder, multimodal fusion, and decoder components. This enhanced model aims to facilitate more effective communication by leveraging multiple modes of information. The encoder module in our proposed model is designed to aggregate external knowledge by utilizing a series of transformer blocks. These blocks enable the integration of diverse information sources, thereby enriching the context for generating meaningful dialogue responses. To select the most appropriate response, we treat the optimal response as the ground truth and input it into a separate transformer block, which further refines the model's understanding. For multimodal fusion, we employ a coattention mechanism that effectively aggregates the encoding features derived from the aforementioned encoder module. By combining these features, the model is better equipped to process and integrate various modal data. Subsequently, these aggregated features are connected to a transformer-based decoder, which is responsible for generating the final dialogue response. Through the incorporation of an encoder, multimodal fusion, and decoder, our proposed model offers a more comprehensive and academically rigorous approach to dialogue generation, allowing for richer context and enhanced understanding of the conversation at hand.

We implement a prototype of our proposed human-machine dialogue system and assess its performance using part of the OpenViDial dataset [12]. Evaluation results reveal that our system generates emotionally rich responses, yielding a more human-like interaction experience compared to existing methods. Our contributions are threefold:

(i) We develop an innovative and emotionally intelligent dialogue system that integrates multimodal data collection and processing, emotion-aware response generation, and immersive digital human interactions.

(ii) We introduce a novel emotion-aware multimodal generation model that employs a coattention mechanism for the encoding and fusion of various inputs and a transformer block for decoding and generating responses.

(iii) Our implementation and evaluation of an open dialogue dataset illustrate the improvements in generating high-quality responses with a better user experience.

2. Human-Machine Dialogue System

In this section, we comprehensively describe the newly designed human-machine dialogue system, which incorporates multiple dimensions of information, including traditional text-based knowledge, emotional information, and additional semantic insights gleaned from images. This integration aims to enhance the system's ability to generate dialogues that more closely resemble natural, human-like conversations. By incorporating these diverse sources of information, the dialogue system is better equipped to comprehend and respond to complex conversational contexts, thereby improving the overall user experience.

2.1. System Overview. We present a comprehensive system overview, as depicted in Figure 2, which encompasses three fundamental modules: data collection and processing, dialogue engine, and interaction module. The primary function of the *data collection and processing module* is to obtain raw multimodal inputs from users, such as vocal utterances (then extracting text and emotion from it) and real-time visual images. The module employs an automatic speech recognition model to extract raw text input from the user's voice. Moreover, we adopt the facial recognition network outlined in [13] and subsequently develop an expression recognition network to determine the user's emotional state from the given input image. It is worth noting that vocal intonations are significant carriers of emotion. Consequently, we employ an RNN-based recognition approach to decode the emotional content of the user's voice. Specifically, emotions are classified into five primary categories: happiness, sadness, anger, surprise, and neutral.

Subsequently, we develop the *core dialogue engine*, which comprises an encoder, multimodal fusion, and a decoder to produce a coherent and contextually appropriate response. To enrich the system's semantic knowledge base and facilitate the generation of diverse, engaging, and influential responses, we incorporate an external knowledge resource, thereby avoiding repetitive and monotonous answers.

Finally, the *interaction module* employs a text-to-speech (TTS) technique to convert generated textual responses into realistic and natural-sounding voices. This module encompasses TTS processing and incorporates digital human driving and rendering techniques to establish a well-experienced digital human service system.

2.2. Data Collection and Processing. This module collects and processes the captured data, such as facial and emotional recognition and voice-to-text conversion. By integrating insights from these diverse modalities, our system deeply comprehends the user's emotions and intentions. We next introduce the details of data collection and processing.

2.2.1. Face Recognition. For our robust facial recognition capabilities, we harnessed the power of a groundbreaking single-shot object detection model, YOLO ("You Only Look Once") [13]. YOLO distinctively deviates from the

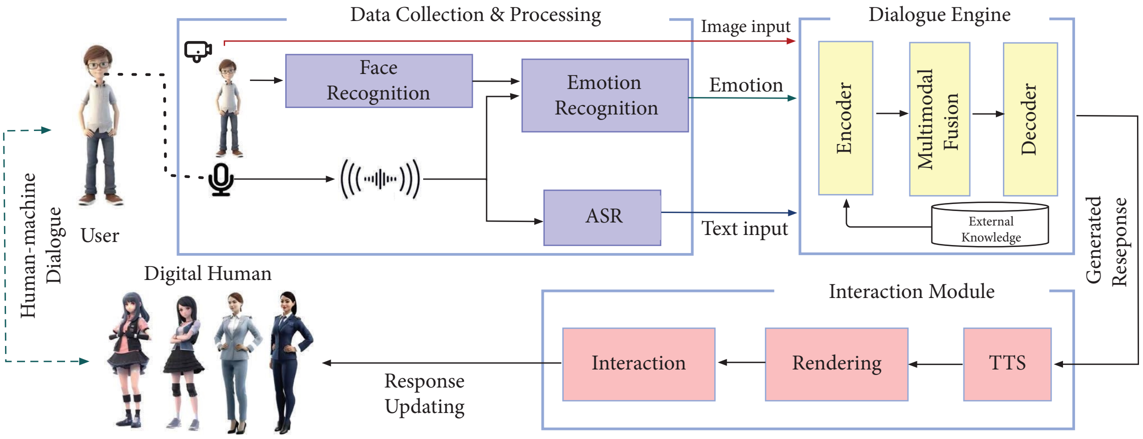

Figure 2: Overview of the proposed multimodal dialogue system. The character images in the figure are created by us using the Midjourney robot (https://discord.com/invite/midjourney), and the rest of the constructions are manually crafted originals.

constraints of conventional region-based detectors, pioneering a grid-based detection strategy. In this schema, images are systematically segmented into cells, where each cell shoulders the responsibility of detecting objects with centers located within its precincts. This ensures an optimized outcome by negating the effects of overlapping or nonpreferential bounding boxes. In tailoring our solution, we employed the avant-garde YOLO Ultralytics iteration, specifically the YOLOv5 model. Our affinity for this choice hinged on its adaptability, its streamlined design encompassing a mere 7.3 million parameters, and its impeccable integration prowess. Our attention was particularly drawn to the YOLOv5's nimble version, dubbed YOLOv5s. This iteration, grounded in the PyTorch framework, heralds innovations such as autonomous learning bounding box anchors. Setting it apart from earlier versions, YOLOv5 is revered for its resource-savvy architecture and sophisticated features. Collectively, these attributes have been instrumental in bolstering the precision and effectiveness of our facial recognition system.

2.2.2. Emotion Recognition. We designed recognition models dedicated to efficiently processing and discerning emotion-relevant data from each modality: an image-centric recognition model and a speech-centric one. Recognizing the distinct natures of visual and auditory cues in expressing emotions, we treated each modality separately. For image-based emotion detection, we chose a convolutional neural network (CNN) architecture, celebrated for its ability to grasp spatial patterns and layered features, rendering it apt for distinguishing nuanced facial expressions and emotion-evoking visual cues. Conversely, the speech-based model employed a recurrent neural network (RNN) framework, capitalizing on its prowess in identifying sequential dependencies and the temporal intricacies within speech patterns. Given that emotions often reveal themselves in vocal variations and tonal shifts, RNNs shine in recognizing these detailed temporal nuances. Through the strategic employment of CNNs and RNNs, we optimized the extraction of each modality's distinct emotional signals. We then fuse the recognition results based on CNNs and RNNs by means of average weighting, leading to enhanced reliability and precision in emotion detection.

2.2.3. ASR. Then, to convert speech signals into text, we have employed a state-of-the-art combination of technologies, namely, the HMM-DNN (Hidden Markov Model-Deep Neural Network) framework and TDNN + LSTM (Time Delay Neural Network and Long Short-Term Memory) networks. The HMM-DNN framework is a harmonious blend of traditional Hidden Markov Models and deep neural networks. While the Hidden Markov Model is tailored to delineate acoustic features, thereby capturing the intricate state transition relationships within sequences, the deep neural network broadens its horizon to extract abstract, high-level features. This symbiotic integration amplifies the robustness of both methods in speech recognition, significantly enhancing model fidelity and precision. Furthermore, our choice of TDNN + LSTM networks strategically merges the strengths of TDNN and LSTM networks. While TDNN shines in the local modeling of speech signals, LSTM excels in recognizing temporal intricacies and spanning long-term dependencies. This fusion ensures that speech features are modeled with heightened efficacy, encapsulating speech content nuances across diverse time frames.

2.3. Methods of Dialogue Engine. We introduce the design of the dialogue engine in Figure 3, consisting of three primary components: an encoder, a multimodal fusion module, and a decoder for response generation. Besides, to improve the quality of generated response, we augmented the knowledge scale by integrating an external knowledge graph. Specifically, the encoder is designed to accept three types of input data: text (derived from audio recognition), emotions (identified from the audio and user visions), and captured user visions. These inputs are subsequently processed through layers, namely, BERT [14], ResNet [15], graph attention [16], and transformer [3], for feature extraction. Notably, the input text is integrated into an external

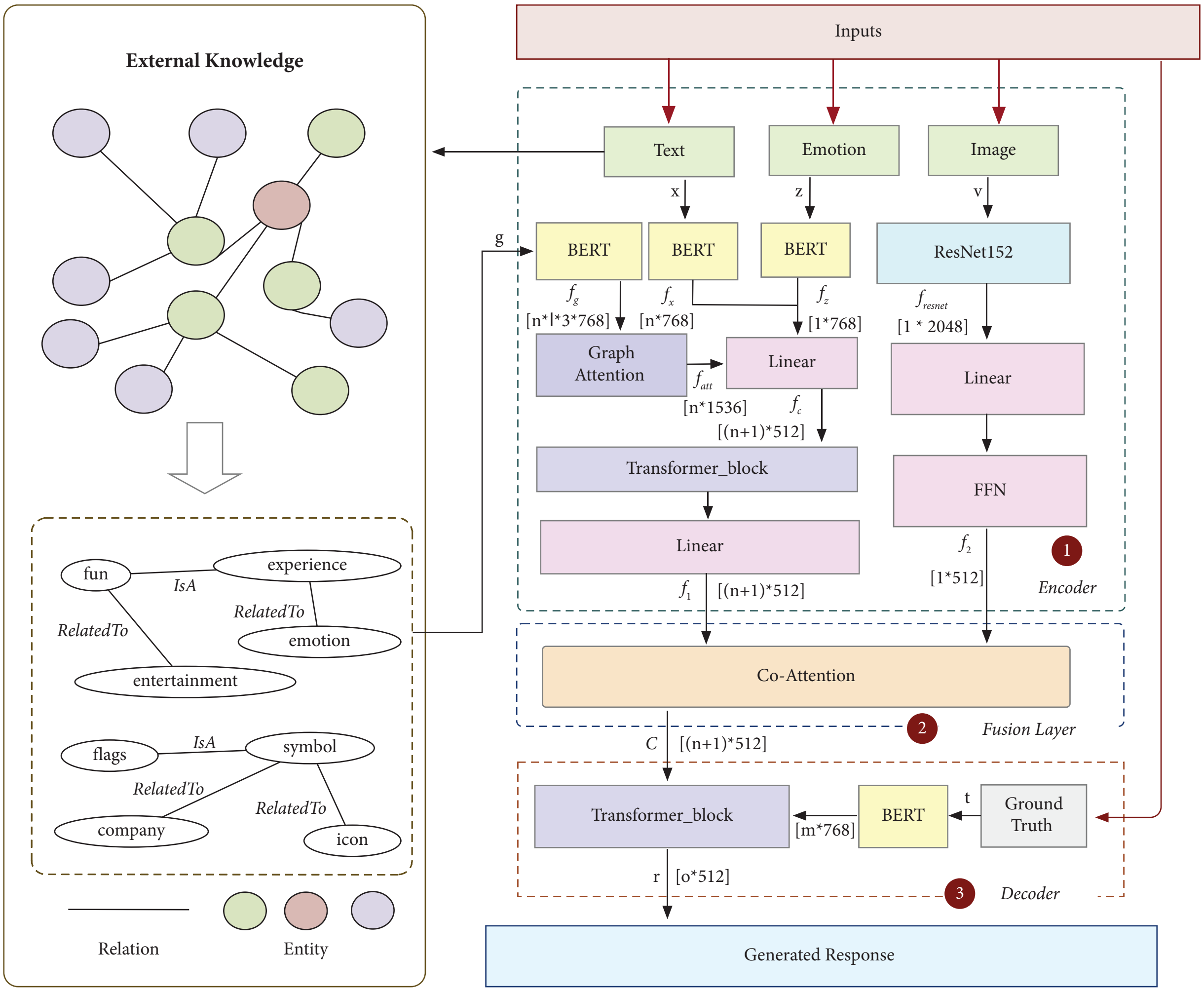

Figure 3: Design of the dialogue engine, an end-to-end multimodal input fusion generation structure based on the transformer, comprising an encoder, a fusion layer, and a decoder (original figure created by us).

knowledge graph, extracting affiliated entities and their semantic associations to bolster the fluency and logic of the generated response. The external knowledge utilized in the encoder is represented as triplets, in the form of ⟨head_entity, relationship, tail_entity⟩, where two entities are interconnected via a relationship. This external knowledge graph is retrieved from a large-scale knowledge base along with the input text. For instance, when the input text is "Welcome to Fun with Flags!," entities such as "fun" and "flags" (as seen in Figure 3) are identified as entities of the knowledge base; subsequently, they are centralized for retrieval. For common words (e.g., "to") that do not have corresponding entities in the knowledge base, a knowledge subgraph containing a special symbol "Not A Fact" is employed. A range of entities associated with them is extracted and then fed into a BERT structure to be encoded in conjunction with other encoding vectors. Subsequently, the encoder outputs the encoding results through linear and FFN layers, serving as input to the feature vector fusion layer based on coattention [17] mechanism. Regarding the decoder design, the merged output is relayed into a transformer-block layer. Ground truth plays a pivotal role exclusively during the training phase, where its primary function is to steer the model in generating accurate responses. This crucial reference acts as a standard, assisting the model in comprehending the intended output patterns. However, it is crucial to emphasize that during the inference phase, the model operates independently without relying on the scaffold of ground truth.

The objective of the aforementioned end-to-end multimodal dialogue generation model is to train and derive responses that are fluent, consistent, and diverse while simultaneously aligning with the current user's textual input and emotional state. Specifically, the input of the model consists of the user text denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, emotion $\mathbf{z}$, introduced external knowledge represented as $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, and image $\mathbf{v}$. Here, $\mathbf{x}$ represents the user text with n words, $\mathbf{g}$ denotes external knowledge, and the knowledge associated with each text consists of n knowledge subgraphs. Each subgraph consists of l triples denoted as $g_i = (k_1, k_2, \ldots, k_l)$. Each triple (i.e., ⟨head_entity, relation, tail_entity⟩) is denoted as $\mathbf{k_i} = (h_i, r_i, t_i)$. Let $\mathbf{t} = (t_1, t_2, \ldots, t_m)$ denote the ground truth response, which consists of m words, and let $\mathbf{r}$ denote the generated response of the model. Our goal is to learn a model $\mathscr{M}$ such that the generated response $\mathbf{r}$ is as close as possible to the gold response $\mathbf{t}$, achieved by modeling the generated response as $\mathbf{r} = \mathscr{M}(\mathbf{x}, \mathbf{z}, \mathbf{g}, \mathbf{v})$.

2.3.1. Encoder Design. In the design of the encoder, we provide an in-depth exposition of the encoder's design based on the aforementioned definitions. Initially, for processing the textual and visual modalities, we, respectively, employ BERT and ResNet152. The textual input, denoted as $\mathbf{x}$, along with the recognized emotion $\mathbf{z}$, is first converted into high-dimensional representative feature vectors of 768 dimensions using BERT. This is mathematically articulated as follows:

$$\begin{aligned} f_x &= \mathrm{BERT}(\mathbf{x}), \\ f_z &= \mathrm{BERT}(\mathbf{z}). \end{aligned} \tag{1}$$

Concurrently, the external knowledge graph, based on user text $\mathbf{x}$, undergoes BERT encoding, represented as

$$f_g = \mathrm{BERT}(\mathbf{g}). \tag{2}$$

Subsequently, we further encode these triples within a graph attention structure to acquire high-level semantic feature vectors of 1536 dimensions, f_{att}, depicted as

$$f_{\mathrm{att}} = \mathrm{GraphAttention}\left(f_g\right). \tag{3}$$

The specific calculation of the graph attention for each subgraph is as follows:

$$\begin{aligned} \beta_j &= \left(W_r \cdot f_{\mathrm{r}_j}\right) \tanh\left(W_h \cdot f_{h_j} + W_t \cdot f_{t_j}\right), \\ \alpha_j &= \frac{\exp\left(\beta_j\right)}{\sum_{j=1}^{l} \exp\left(\beta_j\right)}, \\ f_{\mathrm{att}} &= \sum_{j=1}^{l} \alpha_j \left[f_{h_j}; f_{t_j}\right], \end{aligned} \tag{4}$$

where [;] is concatenation and W_r, W_h, and W_t are trainable parameters.

Following this, the encoded vectors, f_{att}, f_x, and f_z, are concatenated. They are processed through a linear layer to get a vector of 512 dimensions, and then the vector is fed to a Transformer_block layer (encompassing a self-attention mechanism and a feed-forward network) and another linear layer of dimension 512 to yield the encoded result for the textual modality, f_1. This transformation is represented as

$$\begin{aligned} f_c &= \left[f_{\mathrm{att}}; f_x; f_z\right], \\ f_1 &= \mathrm{Linear}\left(T_{\mathrm{block}}\left(\mathrm{Linear}\left(f_c\right)\right)\right). \end{aligned} \tag{5}$$

For encoding the visual input, we utilize the ResNet152 pretrained network as our backbone structure, known for its adeptness in multiscale information capture and superior feature extraction capabilities. ResNet152 is a deep residual network consisting of 152 layers and is pretrained on the ImageNet dataset [18]. We obtain the implementation directly from Torchvision [19] and fine-tune the obtained pretrained model to extract deep-level and high-dimensional features. We extract the 2048-dimensional high-level semantic features generated by the final pooling layer of ResNet152. Subsequently, these refined features are then fed into a linear layer and a single-layer feed-forward network (FFN) to produce the final feature representation f_2 for visual image $\mathbf{v}$. The FFN is a multilayer perceptron with ReLU activation units and a final layer of 512 dimensions. Hence, this methodology enhances the efficiency and accuracy of our dialogue engine, allowing for the effective integration of diverse input modalities. We describe the visual modal input encoding process as follows:

$$\begin{aligned} f_{\mathrm{resnet}} &= \mathrm{ResNet152}(\mathbf{v}), \\ f_2 &= \mathrm{FFN}\left(\mathrm{Linear}\left(f_{\mathrm{resnet}}\right)\right). \end{aligned} \tag{6}$$

2.3.2. Multimodal Fusion Layer. To facilitate better mutual understanding between textual and visual modal information, as well as to extract rich semantic knowledge, we employ a coattention mechanism. This mechanism aims to meld the semantics of both these modalities and derive a representation that captures the codependencies of text and images. In simpler terms, it aims to understand how a piece of text might relate to a visual component and vice versa. We begin with two representations: f_1, a vector for textual data, and f_2, a feature representation of the visual image. The textual vector f_1 is processed through an LSTM network, which is a recurrent neural network (RNN) that excels at managing sequences and capturing long-range dependencies:

$$H_1 = \mathrm{LSTM}\left(f_1\right). \tag{7}$$

For the visual representation, f_2, a simple linear transformation is applied:

$$H_2 = \mathrm{Linear}\left(f_2\right). \tag{8}$$

To capture more intricate relationships, we apply the tanh activation function to the LSTM's output H_1:

$$H_1 = \tanh\left(W_1 \cdot H_1 + b_1\right). \tag{9}$$

We then aim to measure the interaction between the transformed text and image vectors. This is done using the correlation matrix L, which depicts how each segment of H_2 (image) corresponds with every section of H_1 (text):

$$L = \left(H_2\right)^T \cdot H_1. \tag{10}$$

The following phase computes the attention weights for the textual and visual modalities. These weights, AW_1 and AW_2, spotlight the regions of text and visual image that are most contextually linked:

$$\begin{aligned} \mathrm{AW}_1 &= \mathrm{softmax}(L), \\ \mathrm{AW}_2 &= \mathrm{softmax}\left(L^T\right). \end{aligned} \tag{11}$$

With these attention matrices in hand, we derive context-aware representations. C_1 symbolizes the text-aware image representation, and C_2 stands for the image-aware text representation. Both reflect the original data, but now they are tinged with the context of their counterpart:

$$\begin{aligned} C_1 &= H_2 \cdot \mathrm{AW}_2, \\ C_2 &= [H_1; C_1] \cdot \mathrm{AW}_2. \end{aligned} \tag{12}$$

We then combine these representations, C_1 and C_2, and transform them linearly using a 512-dimensional linear layer to produce a robust representation ready for the decoding process:

$$C = \mathrm{Linear}([C_1; C_2]). \tag{13}$$

In essence, this multimodal fusion layer serves as a bridge. It merges text and images, producing a unified representation that is abundant in both linguistic and visual information, all set for subsequent processing.

2.3.3. Decoder and Training. In this section, we dive into the decoding process, specifically how the feature vector is processed and utilized for prediction. Once the fused attention feature vector C is acquired, it undergoes a transformation via the transformer structure. In essence, the transformer is a cutting-edge architecture predominantly used for sequence-to-sequence tasks. Its utility in our process is to generate a contextually relevant embedding, denoted by $h(C)$. Mathematically:

$$h(C) = T_{\mathrm{block}}(C), \tag{14}$$

and having this transformed vector, it is paramount to convert it into a probability distribution to aid predictions. This is achieved using the softmax function. Softmax is a vital function in neural network architectures that squashes its inputs, typically termed as logits or scores, into a range between 0 and 1, such that they can be interpreted as probabilities. Given the transformed vector, $h(C)$, the softmax is applied in the following manner:

$$P(y_t | y_{<t}, \mathbf{x}, \mathbf{z}, \mathbf{g}, \mathbf{v}) = \mathrm{softmax}(W \cdot h(C) + b), \tag{15}$$

where W and b are trainable parameters, often referred to as the weights and bias, respectively. Their role is to adapt and optimize during training to allow for accurate predictions. Training of this model is grounded in the standard Seq2Seq (sequence-to-sequence) approach. This methodology is largely used in tasks like machine translation or any application where an input sequence needs to be transformed into an output sequence. The objective during training is to minimize the difference, or error, between the predicted sequence and the actual ground truth sequence. This is quantified using the negative log-likelihood of producing the true target sequence, represented as

$$\mathrm{Loss} = \sum_{t=1}^{n} -\log(P(y_t | y_{<t}, \mathbf{x}, \mathbf{z}, \mathbf{g}, \mathbf{v})). \tag{16}$$

The above equation calculates the cumulative error across all predictions, aiming to reduce this value during training. For a more comprehensive dive into the training intricacies and the implementation specifics of the dialogue engine, refer to the dedicated implementation.

2.3.4. Analysis of Model Complexity. In comparison to established dialogue generation systems, the complexity of our designed model remains consistent in terms of network architecture and parameter count. Notably, while keeping complexity unchanged, our model introduces a seamless integration of multiple modalities, encompassing text, emotions, and visual inputs. This integration is largely facilitated by our specially designed fusion layer. Upon evaluation, the parameters and complexity of this fusion layer align closely with advanced model structures currently in the field. Consequently, even with no added complexity, our system exhibits marked improvements in the quality, coherence, and fluency of generated dialogues, distinguishing itself from conventional dialogue systems.

2.4. Interaction Module. The interaction module forms the nexus between the user and the system, designed specifically as a Flask-based web application. This choice allows us to leverage Flask's lightweight nature and its ability to quickly spin up web services, giving users a seamless and intuitive interface to interact with the underlying system. Central to our module's capabilities is its text-to-speech (TTS) functionality. Here, we employ two state-of-the-art technologies: DIAN (Deep Iterative and Adaptive Network) [20] and LPCNet (Linear Predictive Coding Network) [21]. These technologies have carved a niche in the TTS realm due to their unparalleled proficiency in generating digital human voice outputs that resonate closely with a natural human voice. This essentially means that users can expect a lifelike, authentic auditory experience, negating the robotic tone often associated with older TTS systems.

On the visual front, the digital representation of the human is not a mere static image. Instead, we bring it to life using a sophisticated 3D mesh. This intricate design approach, coupled with advanced digital human rendering techniques, ensures the resultant visuals are not just realistic but also deeply engaging. The lifelike visuals effectively bridge the uncanny valley, providing users with a more immersive interaction. Understanding the evolving needs of modern applications, we have taken steps to ensure our system is not just standalone. To this end, we have developed a robust set of application programming interfaces (APIs). These APIs serve multiple purposes, from the management of the system and scheduling services to enabling seamless integration capabilities with other existing systems. This modular approach provides flexibility, allowing businesses or developers to tailor the interaction module to fit varied application contexts. Lastly, when it comes to the delivery of inference results, we have ensured that they are not just presented in a linear or one-dimensional manner. Instead, our outputs are structured to offer a holistic, multimodal user interaction service. This is pivotal as it recognizes and respects the varied interaction preferences of users. Whether one leans more towards auditory, visual, or a mix of interaction modes, our system stands ready to cater to their unique needs.

3. Evaluation

3.1. Settings and Datasets

3.1.1. Settings. We will present the key implementations and parameter settings for our dialogue system in a step-by-step manner.

(i) *Emotion Recognition Implementation. (1) Architecture Choice.* For the task of emotion recognition, we utilized multiple layers of RNNs, which are specifically designed to capture the intricate time-dependent and emotional nuances present within speech signals. *(2) RNN Configuration.* Our RNN setup comprises three layers. Each of these layers contains 128 hidden units. *(3) Deep CNN Integration.* Alongside RNNs, we have integrated a deep CNN setup. This architecture features five convolutional layers. Each layer varies in terms of the number of filters, spanning a range from 32 to 128.

(ii) *Dialogue Engine Implementation. (1) Platform.*We built our dialogue engine using the PyTorch deep learning library, known for its versatility and efficiency in handling sophisticated neural network designs. *(2) Model Construction.* Our dialogue model leans on the strengths of fine-tuned BERT and ResNet152 models. We use the "*BERT-Base*" pretrained model released by Google, which consists of 12 transformer encoder layers, 12 multiattention heads, a hidden layer size of 768, and a parameter count of 110 million. The ResNet152 model comprises 60.19 million parameters. This ensures that the generated dialogues are not only meaningful but also contextually relevant. *(3) Encoder Configuration.* The transformer architecture in the encoder has 4 layers, each containing 512 hidden units and 6 attention heads. The linear layer within the encoder maintains a vector dimension of 512. *(4) Decoding Strategy.* The decoding process also employs a transformer architecture. The specifications for this transformer include 4 layers, 512 hidden units, and six attention heads. *(5) Training Settings.* During training, we harnessed the Adam optimizer to refine our model. The chosen batch size is 32, coupled with a learning rate set at 0.0001. To combat overfitting, we introduced L2 regularization, which penalizes the model's weights. Our selection of these hyperparameters is backed by thorough testing and a comprehensive review of existing literature. *(6) Inference Strategy.* When it comes to generating responses, we employ a beam search approach. The set size for this strategy is 2, ensuring the output responses are both varied and of premium quality.

3.1.2. Datasets. We delineate the datasets utilized for different models within our proposed dialogue system and elaborate on the associated data preprocessing techniques.

IEMOCAP Dataset for Emotion Recognition. For our study in emotion recognition, we selected the IEMOCAP dataset [22]. IEMOCAP is notable for its genuine and spontaneous interpersonal interactions that span a wide range of emotional expressions. It encompasses five primary emotional categories, namely, anger, happiness, neutrality, sadness, and surprise. The dataset, consisting of over 12,000 instances, has been partitioned into training, validation, and test sets, adhering to an 80%/10%/10% split.

OpenViDial Dataset for Dialogue Generation. To assess the effectiveness of our dialogue generation model, we utilized a subset of the OpenViDial corpus [12], containing 80,000 single-round open-domain dialogues. Each sentence in this dataset is juxtaposed with a corresponding visual context.

(i) *Emotion Labeling.* We attributed an emotion label—happiness, sadness, anger, surprise, or neutral—to each instance within this dataset.

(ii) *Knowledge Extraction from ConceptNet.*Leveraging ConceptNet [23], an external knowledge base, we extracted knowledge triples for each dataset item. This extraction involved fuzzy matching of word forms in the input text with Spacy (https://spacy.io/). Prioritizing verbs and nouns as potential knowledge concepts, we discarded stop words. Using a preloaded language model, we discerned and segregated verbs and nouns for further analysis. The segmented word results were then screened to omit stop words. This methodology, which echoes the approach of Guan et al. [24], ensures that our extracted graph remains uncluttered.

(iii) *Subgraph Creation.* After identification of primary concepts, we delved into the knowledge base to explore and extract neighboring concepts. This formed a 1-hop knowledge subgraph. Though this process could be iteratively performed for multihop knowledge extraction, we confined our exploration to 1-hop for training efficiency. We also limited the neighboring concept count to a maximum of 100 per primary concept.

In the end, we structured the dataset into three segments: a test set with 8,000 samples, a validation set also with 8,000 samples, and a training set containing the remaining 64,000 samples.

3.1.3. Baselines. To evaluate the effectiveness and usability of our proposed multimodal and emotion-aware dialogue system, we compared it against leading unimodal and multimodal dialogue systems. The systems we considered include

(i) *Text-Based Method.* This advanced dialogue generation method is based solely on text input, as cited in [25]. We use the architecture of the generation stage.

(ii) *Emotion-Based Method.* This dialogue system is cognizant of the user's emotions. It derives its responses by focusing on the user input text and the displayed emotions while excluding any visual cues from the user, as referenced in [6].

(iii) *Image-Chat* [4]. This represents a dialogue generation technique that merges both text and visual inputs. The model is equipped with a comprehensive input array consisting of user text, image style, and images. Utilizing all these streams of information allows for a thorough assessment of the model's performance.

Besides, we have labeled our own methods as "Ours-Emotion" (which utilizes only text and emotional inputs), "Ours-Visual" (which draws upon text and visual data), and "Ours" (which incorporates text, emotion, and visual inputs) to denote the different configurations and their corresponding generated results.

3.2. Overall Performance

3.2.1. Automatic Evaluation. To evaluate the performance of our system automatically, we make use of eight well-established metrics: Bleu [26], Nist [27], Rouge [28], Meteor [29], Diversity [30], and Informativeness [31]. These metrics are designed to quantitatively assess the quality of the generated text in comparison with the target text. To assess the quantitative results more comprehensively, we introduced two metrics from [32], Context Coherence and Language Fluency. Context Coherence assesses the coherence between the input text and the generated response. Language Fluency is based on the negative perplexity of generated response. We then introduce the metrics for evaluating the performance of the proposed dialogue engine.

(i) Bleu (Bilingual Evaluation Understudy) score is a metric that calculates the degree of similarity between the generated sequences and the reference sequences by examining the cooccurrence of n-grams of varying lengths (1, 2, and 4). A higher Bleu score indicates a greater overlap and, hence, a better agreement between the generated and reference texts.

(ii) Nist (National Institute of Standards and Technology) score is an improvement over Bleu in machine translation by introducing the concept of the information value of each n-gram, which assigns a higher weight to keywords that appear less frequently.

(iii) Rouge (Recall-Oriented Understudy for Gisting Evaluation) score is a recall-based measure that computes the number of overlapping n-grams of lengths 1, 2, and 4 between the generated and reference responses. In our analysis, we focus on the F-scores of Rouge, which is a harmonic mean of precision and recall. This measure offers valuable insights into the proportion of relevant information present in the reference responses that our model successfully captures.

(iv) Meteor (Evaluation of Translation with Explicit Ordering) metric compares the similarity between a reference response and a generated response by mapping them to a common space. It considers various factors such as word and phrase matching, word order, grammar, and semantics.

(v) Diversity metric evaluates the richness of the generated response by counting the number of distinct n-grams of lengths 1, 2, 3, and 4 in the generated responses. To ensure a fair comparison, the Diversity metric is scaled by the total number of generated tokens, thereby accounting for potential differences in sentence length. A higher Diversity score implies a more varied and creative output.

(vi) Word level Entropy (Ent-4) is a metric that can be used to evaluate the amount of information generated in text generation. It measures the uncertainty or unpredictability of the next word in a sequence. A higher Ent-4 value indicates higher uncertainty and, therefore, more information.

(vii) Context Coherence refers to the degree of coherence or consistency in the context of a generation model. It measures how well the generated response flows and maintains consistency with the given context.

(viii) Language Fluency refers to the ability of a generated response to sound natural and fluent as if it were written or spoken by a human. It includes aspects such as grammar, syntax, vocabulary, and style. Fluency is an important aspect of text generation, as it affects how easily the response can be read and understood by human readers.

By employing these eight metrics, we can effectively assess the overall performance of our system in terms of both its similarity to the reference responses and the diversity of its generated output.

Table 1 offers a comprehensive, objective comparison across different dialogue models, each encapsulating a different fusion of modalities. A closer inspection offers the following insights:

(i) *Multimodal Advantage*. The superior performance of the "Ours" model, as evinced by the highest scores across all metrics, emphasizes the cumulative benefits of integrating textual, visual, and emotional data sources. This observation aligns with current academic perspectives advocating for the fusion of different modalities to better understand user inputs in conversational AI systems. Interestingly, "Image-Chat," another multimodal competitor, showcases the commendable performance, underscoring the growing consensus in the academic community about the power of multimodality. However, its slight underperformance compared to "Ours" accentuates the pivotal role emotions play in enhancing dialogue quality.

Table 1: Objective experimental results of different methods (higher values indicating better models).

Model	Bleu-2	Nist	Rouge-L	Meteor	Dist-2	Ent-4	Coherence	Fluency
Text-based	0.2921	2.884	0.0157	0.0098	0.0069	1.327	0.1906	0.1697
Emotion-based	0.303	2.8012	0.0169	0.0101	0.0082	1.379	0.2089	0.1731
Image-Chat	0.3313	3.0531	0.0199	0.0128	0.01	1.621	0.2311	0.2097
Ours-Emotion	0.3055	2.9805	0.0168	0.0102	0.0077	1.38	0.1909	0.1718
Ours-Visual	0.3243	3.0104	0.0188	0.0116	0.0093	1.67	0.2038	0.1879
Ours	0.3665	3.2483	0.0225	0.0147	0.0112	1.86	0.2696	0.2306

(ii) *Evaluation of Modalities*. "Emotion-based" results have marginal performance gains compared to "Text-based" results, such as Blue-2 and Meteor. However, "Ours" and "Image-Chat" have a great effect improvement, which indicates that the input of multiple modalities plays a more important role in improving the objective metrics of all aspects of the model. Furthermore, when compared with "Ours-Emotion" and "Ours-Visual," this further confirms the significance of utilizing multiple modalities.

(iii) *Linguistic Quality and Relevance*. The "Ours" model's Meteor, Rouge-L, and Bleu-2 scores emphasize its capability to generate responses that are linguistically aligned with reference responses. A Meteor score of 0.0147, significantly higher than other models, highlights the model's ability to map generated responses to a reference space considering word order, semantics, and syntax. This underscores the model's strength in preserving linguistic quality. The Rouge-L F-scores, which focus on the harmonic mean of precision and recall, underscore the "Ours" model's capacity to retain salient information from reference responses. This is indicative of the model's prowess in generating responses that do not stray from the expected context.

(iv) *Diversity and Informativeness*. In generating human-like responses, the variance in responses is crucial. A monotonous, patterned response can be easily detected by human users and might detract from the user experience. The Diversity metric (Dist-2) and the Word level Entropy (Ent-4) of "Ours" show the model's potential to deliver varied yet contextually appropriate responses. This emphasizes the model's ability to navigate the trade-off between randomness and relevance, a feat often challenging in NLP applications.

Moreover, the findings highlight several avenues for future research:

(i) *Emotion Utilization*. Given the nuanced performance enhancements between "Text-based" and "Emotion-based," there is a clear need for further studies examining how best to leverage emotional cues in dialogue systems. This would pave the way for more emotionally intelligent AI systems, which is crucial for certain application areas like mental health or customer service.

(ii) *Metric Evolution*. As the complexity and richness of dialogue models grow, there is a pressing need for the evolution of metrics that can holistically capture the performance nuances. Traditional metrics, while valuable, might not fully encapsulate the breadth of capabilities of advanced multimodal systems.

3.2.2. Manual Annotation. For the manual evaluation, we systematically selected 100 random cases from the test set for each model and enlisted five volunteers to participate in a blind test comparison between the model-generated results and the ground truth responses. The volunteers were instructed to rate the "more appealing" response out of two options, one being the human-generated response (the ground truth) and the other from the model. The subjective metrics employed for evaluation included the appropriateness of the responses in terms of topic relevance, logical coherence, linguistic fluency, diversity, and informativeness, as well as the emotionality of the responses.

Table 2 showcases the results of manual annotations on the performance of various models concerning subjective metrics. These metrics, designed to gauge user perception and experience, reflect on how these models compare with ground truth responses. Here is an in-depth analysis of these findings:

(i) *Overall Performance*. The "Ours" model consistently outperforms the other models in almost all the subjective metrics with scores of 54.3%, 53.4%, and 46.7% for Appropriateness, Informativeness, and Emotional aspects, respectively. It is worth noting that a lower value indicates a better model. The "Image-Chat" model, despite being multimodal, significantly trails the "Ours" model and even some unimodal models in all metrics. This result possibly underlines the challenge of optimally exploiting the fusion of different modalities.

(ii) *Appropriateness*. The "Emotion-based" model scores 54.2%, reflecting its strength in generating contextually appropriate responses. This result is expected as emotional information inherently anchors the generated response within a context, thereby making it more relevant. The "Text-based" model's score of 57.3% suggests its proficiency in generating topic-relevant responses. Its performance, combined with the higher performance of the "Image-Chat" model (58.2%), indicates that textual context, in some cases, might be more

Table 2: Manual annotation results of different methods (lower values indicating better models).

Ground truth vs	Appropriateness (%)	Informativeness (%)	Emotional (%)
Text-based	57.3	69.3	59.8
Emotion-based	**54.2**	62.7	54.2
Image-Chat	58.2	55.1	52.8
Ours-Emotion	63.2	61.8	52.1
Ours-Visual	60.4	60.5	56.5
Ours	54.3	**53.4**	**46.7**

Bold values indicate the best results.

pivotal in ensuring appropriateness than visual information.

(iii) *Informativeness*. "Image-Chat" and "Ours" are far ahead in terms of the amount of information, indicating that visual information plays a great role in the process of dialogue generation. Images provide a visual context, and a single image can encapsulate a complex scene, environment, or emotion, instantly setting the stage for a more meaningful and relevant conversation.

(iv) *Emotionality*. The "Ours-Emotion" model has a higher emotional score, while the "Ours-Visual" model has a lower emotional score. This indicates that the utilization of emotional information significantly impacts the performance of the model in the process of dialogue generation. Emotional information can imbue conversations with more human-like qualities. When comparing the "Text-based" model with the "Emotion-based" model, the "Emotion-based" model shows a notably lower score (54.2%). This further confirms the crucial role of emotional information in dialogue generation. These two observations underscore the importance of emotional information in creating engaging conversations and effectively interacting with humans.

Besides, we conclude some implications from manual annotations.

(i) *Perception vs. Objectivity*. The divergence in performance between manual annotations and objective metrics (as discussed in previous results) accentuates the need for harmonizing these evaluations. The user-centric perception sometimes might not align with objective standards, underscoring the challenge of creating universally acclaimed models.

(ii) *Optimizing Modalities*. The results highlight an opportunity to delve deeper into the nuances of modality interactions. For example, how does the inclusion of visual cues influence perceived appropriateness or emotionality? This insight can pave the way for more user-aligned model optimizations.

(iii) *Emotional Resonance*. The variation in emotional scores across models indicates a research avenue in understanding the dynamics of emotional resonance in AI-human interactions. A granular breakdown of emotional responses (e.g., joy, sadness, and neutrality) in future evaluations can provide richer insights.

In summation, the manual annotations provide a valuable perspective on how users perceive and value model-generated responses. This user-centric evaluation complements objective metrics, creating a comprehensive evaluation framework for dialogue models.

3.3. In-Depth Analysis. In this section, we discuss the performance of emotion recognition methodologies employed within the dialogue system. The metrics utilized for evaluation include Precision, Recall, and the *F*1 score. A head-to-head comparison was established between an emotion recognition method based on SVM [33] and the method we proposed, which combines CNN and RNN. Their performances under various input modalities were also tested. For the sake of clear distinction: SVM-(Speech), SVM-(Visual), and SVM-(Speech + Visual) represent emotion recognition solely from speech, solely from visual inputs, and from a combined modality of both speech and visual inputs, respectively. Similarly, CNN + RNN-(Speech), CNN + RNN-(Visual), and CNN + RNN-(Speech + Visual) depict the results of our proposed method when different modalities are inputted. The principal findings from this analysis are captured in Table 3.

(i) *SVM vs. CNN + RNN*. It can be clearly seen from Table 3 that the proposed CNN + RNN method outperforms SVM in all modalities, and the *F*1 score achieved by CNN + RNN (Speech + Visual) is the highest at 71.91%. This highlights the potential benefits of leveraging deep learning architectures for emotion recognition tasks.

(ii) *Impact of Input Modalities*. Both SVM and CNN + RNN models demonstrate incremental performance improvements as they shift from singular modalities to combined ones. This suggests that integrating both speech and visual information provides a more holistic and accurate representation of emotional states. Specifically, the *F*1 score for the SVM model jumps from 45.54% in the speech-only mode to 59.53% when both speech and visual inputs are integrated. Similarly, for the CNN + RNN model, the *F*1 score ascends from 65.43% (speech-only) to 71.91% (speech and visual combined).

(iii) *Precision vs. Recall*. Upon careful observation of precision and recall values, particularly with the CNN + RNN model, an overall performance balance

Table 3: In-depth analysis of different emotion recognition methods.

Method	Precision	Recall	$F1$
SVM-(Speech)	45.93	45.16	45.54
SVM-(Visual)	52.01	52.10	52.25
SVM-(Speech + Visual)	59.28	59.79	59.53
CNN + RNN-(Speech)	65.93	64.83	65.43
CNN + RNN-(Visual)	69.23	67.29	68.25
CNN + RNN-(Speech + Visual)	**71.81**	**72.01**	**71.91**

Bold values indicate the best results.

can be observed. For instance, the precision and recall of CNN + RNN-(Speech + Visual) are 71.81% and 72.01%, respectively. This indicates that the model accurately identifies emotions (high precision) and also captures a significant portion of actual existing emotions (high recall).

(iv) *Visual Modalities*. Through data analysis of the experimental results, we found that both SVM-(Visual) and CNN + RNN-(Visual) report higher scores than their results of speech-only. This emphasizes the crucial role of visual information in decoding emotions, as visual information often includes some nonverbal information.

Moreover, the consistent outperformance of the CNN + RNN model suggests the advantages of leveraging more complex architectures for emotion recognition, given the intricate nature of emotions. The incremental benefits witnessed as we shift from singular to combined modalities underline the importance of capturing emotions from diverse sources. It points towards a potential research direction where multimodal integration can be further refined.

3.4. User Experience Analysis. To understand how emotions influence user experience in dialogue systems, we carried out a comprehensive assessment. The overarching goal was to gauge user satisfaction and willingness to use a dialogue system that has integrated emotion recognition. Our experiment comprised two versions of our prototype dialogue system: one embedded with the emotion recognition model ("emotional system") and the other without it ("nonemotional system"). The study engaged 30 volunteers. In our participant selection, we strived for diversity, aiming to mirror the demographic spread of potential real-world users. Volunteers interacted with both versions of the system. To eliminate order bias, which could skew perceptions based on sequence, we randomized the starting system for each participant. After interaction, participants were required to score both versions on two metrics: satisfaction and willingness to use, rated on a scale of 1 to 5. While users were encouraged to focus on system performance, we provided guidelines to prevent nonsystem attributes from influencing their evaluation.

We show the results and findings in Figure 4 as follows:

(i) *Overall User Experience*. A glance at Figure 4 showcases the superiority of the emotional system in terms of user satisfaction and willingness to use. Both metrics received a higher mean opinion score (MOS) for the emotional system.

(ii) *Detailed Insights*. The emotional system not only registered higher scores but was also frequently described as more user-friendly and convenient by the participants. There was a discernible trend among users favoring the emotional system, signaling its heightened appeal and potential recurrent usage.

(iii) *Emotional Nuance*. One standout inference is the potential of emotion recognition in capturing semantic depth and richness in user interactions. Emotions, being pivotal to human experiences, can elevate a system's capability to grasp intricate user sentiments, thereby enriching the dialogue quality.

Moreover, the success of the emotional system, as demonstrated by our user study, holds significant potential for the realm of dialogue system design:

(i) *Emotion as a Keystone*. The study underscores the pivotal role of emotional information. Integrating emotions can drastically uplift the quality of interaction, fostering a more holistic, nuanced, and satisfactory user experience.

(ii) *Future Design Considerations*. Designers and developers of subsequent dialogue systems should heavily weigh the advantages of integrating emotional components. Our findings suggest a clear user preference for systems that can cognize and respond to emotional cues.

4. Related Work

4.1. Intelligent Dialogue System. Dialogue systems, which predominantly utilize text or voice input, have become an integral part of various customer service applications, encompassing retail websites and entertainment services [30]. Some of these systems are equipped with intelligence in the form of digital human avatars [34], specifically designed for domain-specific employee training and interview scenarios [35]. These systems primarily function by converting voice input into text and generating intelligent responses through the application of natural language processing (NLP) methods, such as Seq2Seq-based text generation techniques [2]. Nonetheless, as the input is primarily derived from text with single-modal information, the resulting responses tend to be monotonous and unengaging, leading to a diminished user willingness to engage with them. Intelligent dialogue systems based on twin digital humans [11] possess the capability to recognize users' voices, process them accordingly, and respond to questions or engage in casual conversation, depending on the specific application scenarios, thus delivering a more human-like service experience. Researchers such as Cui et al. [36, 37] have explored multimodal dialogue systems tailored for particular industries, such as fashion and retail. Concurrently, Wang et al. [12, 38] have developed dialogue systems employing various natural language generation models for open-domain dialogues. In summary, these studies primarily focus on enhancing the quality of generated responses in

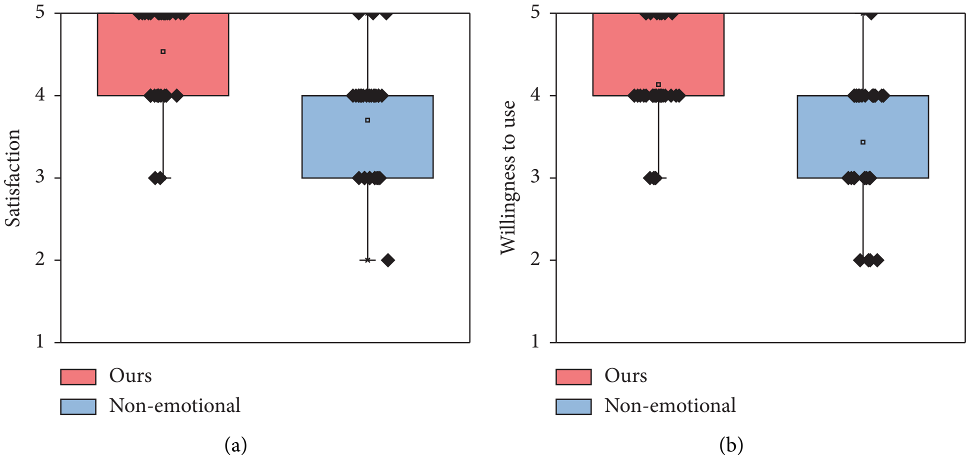

Figure 4: User experience analysis. (a) Results on users' satisfaction. (b) Results on users' willingness to use.

dialogue systems through the application of advanced generation models. However, they generally overlook the importance of incorporating user emotional considerations, particularly in the context of digital service scenarios, which may prove vital in delivering a more engaging and effective user experience.

4.2. Emotional Dialogue System. A considerable amount of research has been conducted on incorporating emotional conversation generation in dialogue systems, with the primary focus on enabling the model to generate emotionally appropriate responses that cater to the user's feelings. This is typically achieved by labeling conversation corpora with emotions and incorporating the emotion labels during the learning process, as demonstrated in the seminal studies conducted by Wei et al. [6] and Li et al. [7], who utilized this strategy to improve the emotional intelligence of their proposed dialogue systems. Huber et al. [9] introduced an innovative approach by proposing an image-based conversational agent that utilized visual emotions, facial expressions, and scene features to determine the user's emotional state. In their work, emotions were classified into two broad categories, namely, positive and negative, to facilitate a simplified emotional understanding. In a distinct study, Tian et al. [10] put forth a multitask learning framework in which tasks such as image sentiment sequential labeling, image sentiment classification, and text generation were learned simultaneously. This was accomplished using a pretrained model specifically designed to generate textual content that effectively captures the user's emotions. With respect to emotional multimodal dialogue systems, Fung et al. [8] developed a virtual interactive dialogue system aimed at collecting user responses and assessing the Myers–Briggs Type Indicator (MBTI) personality of the user, thereby providing a deeper understanding of the user's personality traits. However, it is important to note that the majority of existing sentiment-based research primarily focuses on extracting user emotions from visual cues, such as facial expressions and body language. In reality, both visual and auditory cues, such as the user's voice, can convey emotions, necessitating a more comprehensive approach to emotion recognition. Consequently, the challenge lies in deeply integrating multimodal information and user emotions to enhance the quality and performance of dialogue systems. Our approach attempts to address this challenge by considering multiple modalities, setting it apart from the existing work in the field.

5. Discussion and Conclusion

This work addresses the challenge of generating nonemotional dialogues and ventures into the realm of intelligent and emotionally responsive human-machine systems. Central to our approach is the extraction of diverse emotions from multimodal inputs and their fusion into a coherent semantic expression to generate an empathetic response. To actualize this, the system employs a coattention mechanism coupled with a transformer-based decoder. Our study delved into the potential of an end-to-end generative model, particularly focusing on the exploration of a comprehensive multimodal input. While reinforcement learning has gained traction in dialogue generation, particularly for single-modal text input tasks, the complexity of an emotion-aware multimodal dialogue using reinforcement learning presented challenges, including the dearth of adequate RL environments and the extensive computational resources required. Consequently, this study steered clear of it.

Preliminary evaluations showed significant advancements when contrasted with the state of the art. The results indicate not just technical superiority but also a noticeable preference among users for our emotional multimodal dialogue system. This preference translates to substantial prospective cost savings, especially in sectors like food and social services, by potentially reducing the dependency on large customer service teams. With our work, we have carved a niche in the domain of human-machine interaction, bridging the gap between emotional understanding and machine response. Our findings suggest an economically viable avenue for enhancing communication across various sectors. While our current efforts have been productive, we believe the horizon is vast. A promising direction for future research will be incorporating reinforcement learning to refine our dialogue system further. As we advance, our endeavor will be to deploy the system across diverse settings to gauge its universal applicability and impact.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yaru Zhao was responsible for conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, original draft preparation, and review and editing. Bo Cheng was responsible for conceptualization, resources, project administration, supervision, and review and editing. Yakun Huang was responsible for conceptualization, methodology, supervision, visualization, and review and editing. Zhiguo Wan was responsible for conceptualization and review and editing. All authors reviewed and approved the final version of the manuscript for submission.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under grant no. 2022YFF0902701, the National Natural Science Foundation of China under grant nos. 62202065, U21A20468, 61921003, and 61972043, the Zhejiang Lab under grant no. 2021PD0AB02, and the Fundamental Research Funds for the Central Universities under grant no. 2020XD-A07-1.

References

[1] W. A. Abro, A. Aicher, N. Rach, S. Ultes, W. Minker, and G. Qi, "Natural language understanding for argumentative dialogue systems in the opinion building domain," *Knowledge-Based Systems*, vol. 242, Article ID 108318, 2022.

[2] I. Sutskever, O. Vinyals, and V. L. Quoc, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 3104–3112, Montreal, Quebec, Canada, December 2014.

[3] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, Long Beach, CA, USA, December 2017.

[4] K. Shuster, S. Humeau, B. Antoine, and J. Weston, "Image chat: engaging grounded conversations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2414–2429, Stroudsburg, PA, USA, July 2020.

[5] K. Shuster, E. Michael Smith, D. Ju, and J. Weston, "Multi-modal open-domain dialogue," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4863–4883, Punta Cana Dominican Republic, November 2021.

[6] W. Wei, J. Liu, and X. Mao, "Emotion-aware chat machine: automatic emotional response generation for human-like emotional interaction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1401–1410, Beijing, China, November 2019.

[7] S. Li, F. Shi, and D. Wang, "Emoelicitor: an open domain response generation model with user emotional reaction awareness," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3637–3643, Yokohama, Japan, January 2021.

[8] P. Fung, A. Dey, and F. B. Siddique, "Zara: a virtual interactive dialogue system incorporating emotion, sentiment and personality recognition," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pp. 278–281, System Demonstrations, Yokohama, Japan, December 2016.

[9] B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan, "Emotional dialogue generation using image-grounded language models," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Montreal, QC, Canada, April 2018.

[10] Z. Tian, Z. Wen, Z. Wu, Y. Song, and J. Tang, "Emotion-aware multimodal pre-training for image-grounded emotional response generation," in *Proceedings of the International Conference on Database Systems for Advanced Applications*, pp. 3–19, Tianjin, China, April 2022.

[11] T. Shen, J. Zuo, S. Fan, J. Zhang, and L. Jiang, "Vida-man: visual dialog with digital humans," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2789–2791, Verlagsort, NY, USA, October 2021.

[12] S. Wang, Y. Meng, and X. Sun, "Modeling text-visual mutual dependency for multi-modal dialog generation," 2021, https://arxiv.org/abs/2105.14445.

[13] G. Castellano, B. De Carolis, N. Marvulli, M. Sciancalepore, and G. Vessio, "Real-time age estimation from facial images using yolo and efficientnet," in *International Conference on Computer Analysis of Images and Patterns*, pp. 275–284, Limassol, Cyprus, September 2021.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Stroudsburg, PA, USA, October2019.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, Las Vegas, ND, USA, June 2017.

[16] H. Zhou, T. Young, M. Huang et al., "Commonsense knowledge aware conversation generation with graph attention," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4623–4629, Stockholm, Sweden, July2018.

[17] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, April 2017.

[18] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[19] N. Ketkar, J. Moolayil, N. Ketkar, and J. Moolayil, "Introduction to pytorch," *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*, pp. 27–91, 2021.

[20] Y. Chen, L. Wu, and M. J. Zaki, "Iterative deep graph learning for graph neural networks: better and robust node embeddings," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 19314–19326, Canada, December 2020.

[21] A. Harma, "Linear predictive coding with modified filter structures," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 769–777, 2001.

[22] C. Busso, M. Bulut, C.-C. Lee et al., "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[23] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: an open multilingual graph of general knowledge," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 4444–4451, San Francisco, CA, USA, February 2017.

[24] J. Guan, Y. Wang, and M. Huang, "Story ending generation with incremental encoding and commonsense knowledge," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 6473–6480, Honolulu, HW, USA, February 2019.

[25] H. Song, Y. Wang, W. Zhang, X. Liu, and T. Liu, "Generate, delete and rewrite: a three-stage framework for improving persona consistency of dialogue generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5821–5831, Stroudsburg, PA, USA, July 2020.

[26] J. Li, W. Monroe, and D. Jurafsky, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, Stroudsburg, PA, USA, July 2016.

[27] D. George, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145, Stroudsburg, PA, USA, November 2002.

[28] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proceeding of the Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004*, pp. 74–81, 2004.

[29] S. Banerjee and L. Alon, "Meteor: an automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, Ann Arbor, MG, USA, June 2005.

[30] C.-Y. Li, D. Ortega, and D. Väth, "Adviser: a toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 279–286, Stroudsburg, PA, USA, July 2020.

[31] L. Mou, Y. Song, R. Yan, G. Li, and L. Zhang, "Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation," in *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 3349–3358, Osaka, Japan, December 2016.

[32] B. Pang, E. Nijkamp, and W. Han, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3619–3629, Stroudsburg, PA, USA, July 2020.

[33] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on affective computing*, vol. 4, no. 2, pp. 183–196, 2013.

[34] Y. Sano, C. S. Leow, and S. Iiday, "Spoken dialog training system for customer service improvement," in *Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 403–408, Auckland, New Zealand, December 2020.

[35] J. Chen, J. Sun, and H. Huang, "An open-source dialog system with real-time engagement tracking for job interview training applications," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 10–15, Berlin, Germany, December 2020.

[36] C. Cui, W. Wang, X. Song, M. Huang, and X. S. Xu, "User attention-guided multimodal dialog systems," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 445–454, Paris, France, July 2019.

[37] L. Liao, Y. Ma, X. He, R. Hong, and T.-S. Chua, "Knowledge-aware multimodal dialogue systems," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 801–809, Seoul, Republic of Korea, October 2018.

[38] Q. Sun, Y. Wang, and C. Xu, "Multimodal dialogue response generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 2854–2866, Dublin, Ireland, May 2022.