2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Empathetic Response Generation System: Enhancing Photo Reminiscence Chatbot with Emotional Context Analysis

Alberto Herrera*, Xiaobei Qian*, and Li-Chen Fu†

Abstract— Dementia affects 50 million people worldwide, underscoring the urgent need for effective interventions to enhance their well-being. While reminiscence intervention shows promise, its implementation is hindered by limited human resources, making machine-aided systems a viable automated solution for seamless photo-reminiscence sessions. In this paper, we introduce an empathetic response generation system specifically designed to enhance a question-only photo-reminiscence chatbot, with a focus on improving emotional context understanding and enhancing conversation engagement. We leverage Transformers to encode dialogue history, infer emotional states from user responses, and extract named entities. By combining template-based utterances with a retrieval chatbot, our system generates relevant and empathetic responses to user replies. Our system's effectiveness is validated through human evaluations using a Likert-like scale to assess engagement levels. The results demonstrate that our approach surpasses both the question-only system and other models from existing works, including retrieval and generated models. This highlights our system's potential to enhance interactions and engagement, advancing technology-driven interventions for dementia that improve well-being and quality of life.

I. INTRODUCTION

Dementia is a syndrome that deteriorates cognition, mainly affecting (but not exclusively) elder people. Currently, around 50 million individuals are affected worldwide [1], with the number of cases growing yearly. This syndrome negatively impacts the affected individuals and their relatives' physical, psychological, social, and economic conditions.

Reminiscence therapy is a non-pharmacological intervention that has the potential to improve the psychosocial outcomes of the affected person. It involves stimulating conversations about past events, either by storytelling or with physical prompts, like photographs, to evoke long-term memories. O' Philbin et al. [2] indicated that reminiscence intervention can positively affect moods, decrease depression symptoms, and improve communication and cognition.

When people with dementia practice reminiscence by recalling past experiences, their life quality can be improved. Conversations about photos can facilitate this process, especially if they are provided by the persons themselves or their caregivers, since the events contained in the image are highly relevant to their personal lives. Thus, photo reminiscence and the technologies required to facilitate this process have become essential to tackle this global issue.

An autonomous system that can facilitate a conversational reminiscence session from a photograph can decrease the burden on caregivers and be useful for the greater public in general. The agent must understand the content of the photos and encourage the user to remember his or her past relevant experiences. This can be done by asking questions, listening to the user's replies, and providing an adequate response. If the system only asks questions, the user might feel verbally hostage, or the conversation can feel disjointed [3]. Likewise, if replies are too simple, the system may seem unintelligent or unappealing.

It is a significant challenge to listen to the user and generate appropriate, engaging responses, especially for open-ended questions.

II. RELATED WORK

Traditional reminiscence therapies use prompts and are guided by a caregiver to help recall important events, places, names, and dates. Outcomes tend to be improved when these elements are personalized. Studies like [2], [4], and [5] reported improvements in cognitive function, anxiety, depressive symptoms, self-esteem, life satisfaction, and personal interaction, although not every indicator was shown to be improved significantly in every study. Although some of the effects are small and inconsistent, reminiscence intervention is still helpful.

Works like [6], [7], [8], and [9] utilized robots or tablets as technological devices to interact with people for reminiscence sessions. Among their variety of findings, they showed that communication can be positively affected. However, only a few works have focused on autonomous photo-reminiscence.

A knowledge-driven photo-reminiscence system was implemented in [10] as part of the MARIO project, deployed on Kompaï-2 robots. The images must be tagged manually with the names of persons, places, etc. The system builds a knowledge graph and creates questions using templates, providing hints for wrong answers. However, they were not able to handle open-ended questions. In [11], they tried to address the issue

*These authors contributed equally. †Dr. Li-Chen Fu is the corresponding author. The authors are with the Department of Computer Science and Information Engineering, National Taiwan University, the Department of Electrical Engineering, National Taiwan University, and the Graduate Institute of Networking and Multimedia, National Taiwan University. This work was supported by the National Science and Technology Council of Taiwan, and the Center for Artificial Intelligence & Advanced Robotics, National Taiwan University, under the grant numbers NSTC 112-2634-F-002-003- & NSTC 112-2223-E-002-019-.

by analyzing the sentiment of the response, encouraging users to talk more for positive reactions.

Elisabot was developed in [12], which is a virtual agent using Telegram. Employing a Visual Question Generation module (with Resnet-101 [13] and LSTM), they generated up to 5 questions for a user-provided image. User's replies are forwarded to an independent seq-2-seq [14] model to generate a reply. Open-source datasets like MS COCO by [15] or Persona-chat [16] were used to train the networks. Four users aged 60 or older tested the system and rated it via a questionnaire. Some participants found it easy to use and a bit silly, while others felt it was engaging and challenging.

As our first study on this topic, we used a tablet to visualize pictures and RoBoHoN to ask questions in [17]. With VGG-16 [18] trained on the USED dataset [19], we could identify the event of the picture. A Markov random field-based algorithm, containing common sense knowledge and loopy belief propagation, uses the detected event and the user's reply to determine the next utterance from a question pool. Simple follow-ups can be made via sentiment analysis and the reply's length. Volunteers rated the system on relatedness, appropriateness, and effectiveness. Results indicate that the robot can ask coherent questions. The system was improved in [20], changing the robotic platform to Zenbo Junior, including scene prediction with VGG-16 and Places365 [21], and increasing the number of possible questions to ask. Results from the study show that participants proactively engage with the robot and find the prospect of reminiscence therapy both effective and enjoyable.

Unlike previous studies, our work introduces the following key contributions:

- We adopt a multi-modal approach to infer the user's emotional state by integrating image context, user replies, and optionally, facial expressions (for conversational robots with frontal cameras). We then train a strategy selection model to determine the optimal response style, incorporating both stylistic elements and contextual encoding. This approach markedly enhances the empathetic impact of the system's responses, leading to more nuanced and effective interactions.
- We utilize recognized entities from user utterances to adjust templates, generating more contextually relevant responses, particularly for open-ended questions. This method improves the system's ability to deliver informative and appropriate replies, fostering a more engaging user experience.
- We combine visual features, the proposed question, user utterances, and the desired style to form a comprehensive input for the response ranking module. This integration enables the system to produce highly relevant responses, optimizing the response generation process and boosting user engagement and satisfaction.

To give you a straightforward impression of our work,

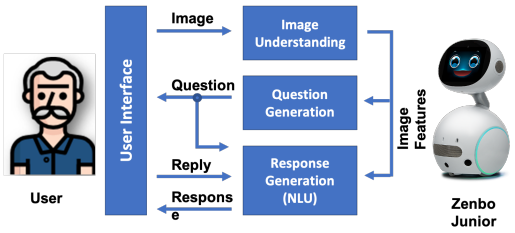

Fig. 1. Photo-Reminiscence System Overview

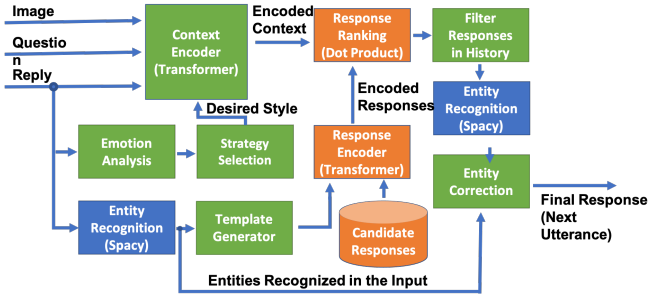

Fig. 2. Response System Overview

we provide a demo video to show the sample conversation between the author and the proposed system concerning two photos. More details can be found on our project website https://ntuairobo.net/remchat_intro.

III. SYSTEM OVERVIEW

Asking questions is a crucial element of a photo-reminiscence system for evoking memories. However, to make the system engaging and useful in the long run, we must keep the conversation human-like. This can be achieved by actively listening to the user and providing appropriate replies to keep the dialogue flowing.

The response generation system presented in this work is meant to enhance the functionality and usability of the photo-reminiscence system presented in [17] and later improved in [20]. Figure 1 shows an overview of how the modules interact, with the proposed one highlighted in green. Users can select images using an Android device and start verbal interaction with the Zenbo Junior robot. The image understanding module will identify the metadata containing the scene and event type from a photograph, which is used by the Question Selection module to ask questions. After the user provides a reply, the robot follows up with one or multiple responses, simulating a conversation between two humans.

An overview of the response generation system can be visualized in Figure 2. The green blocks represent self-created work, orange blocks were adapted from [22], and blue blocks were adapted from [23].

IV. METHODOLOGY

This section describes the functionality details of each block in Figure 2. Starting with the input data (user's

image, user's reply, and previously asked question), the reply is processed by the emotion analysis module, followed by the strategy selection module for style enhancement. Its output is forwarded with the rest of the input data to the context encoder. The reply is also processed by the entity recognition module, whose information is used to create templates. A response encoder module handles these and a pool of static candidates. Once the context and responses are encoded, we rank the utterances via dot product. A final entity correction mechanism is applied as the final step before outputting a response.

A. Emotion Analysis

A system that dynamically adapts the style of the generated replies can provide a more fulfilling experience. We achieve this by detecting the 3-dimensional valence-arousal-dominance(VAD) emotion state of the speaker and shaping the response accordingly. We utilize the model from [24], and our lab's multi-modal model from [25] emotion analysis module. The latter is used only when the robot's frontal camera is available to capture the user's facial expression. This multi-modal model analyzes human multimodal emotions through the proposed disentangled representation learning to enable unimodal capture representations with both modality generality and modality specificity.

In this work, we use the described model as a cloud server to process recorded video input and generate output indicating valence levels. Meanwhile, a video recording service runs in the background during conversations. Specifically, the recorder starts once the robot finishes speaking and continues until the user's voice volume drops below a set threshold for three consecutive seconds, at which point the recording stops.

B. Strategy Selection

To learn the appropriate response strategy, which is a mapping between the user's emotional state and the desired style of response, we utilize Q-learning to train a strategy selection model. We gathering a subset of the Image-Chat dataset for model training; each training sample is a triplet: the first user's utterance with VAD emotional state, the second user's reply with ground-truth style originally labeled in Image-Chat, and the first user's response with VAD emotional state. The model will get a positive reward if the first user's emotion improves, i.e., his level of Valence increases or his level of arousal decreases with consistent negative valence. The Image-Chat [22] dataset divides its 215 style categories into positive, neutral, and negative traits. As the aim of this research work is to encourage users to converse about happy memories instead of sad ones, we condition the output style to only positive ones, with detailed scales along the continuous dimensions of the VAD model.

During the testing phase, upon detecting the VAD emotion state based on the user's reply, we input this information into the strategy selection model to determine the desired style. Integrating the user's emotional state and the response's style into the context enhances the emotional indicators and shapes the form of the response for a more engaging chatbot. This approach ensures that the chatbot's responses are not only contextually relevant but also emotionally appropriate, leading to a more enriching user experience.

C. Entity Recognition and Template Generator

Identifying entities in the user's utterances allows us to provide more targeted responses, making the robot appear more interesting and human-like. Using the names of people or places the user mentions allows us to create more pertinent replies, making our system appear to be better able to understand the user's utterances.

We apply the spaCy Transformers library [23] to mainly detect people and places. A cooldown strategy as inhibition of return is implemented to "forget" entities older than 3 conversation rounds, as they become less relevant for the current topic.

When identifying names, spaCy uses the tag ⟨PERSON⟩ for people and tags like ⟨FAC⟩ (e.g., buildings), ⟨GPE⟩ (e.g., countries, cities), and ⟨LOC⟩ (e.g., mountains) for places. Different templates are created for each label category, following the structure below:

I would love to travel to ⟨GPE⟩ and do this.

⟨GPE⟩ could be replaced by Mexico or Taipei, creating a new utterance. Each template contains a single fillable field. Templates are extracted from the positive utterances of the Image-Chat dataset [22] and are manually analyzed so that those that would not fit well the reminiscence task are removed. There are a total of 179 templates available for the ⟨PERSON⟩ entity type, 62 for ⟨FAC⟩, 168 for ⟨GPE⟩, and 123 for ⟨LOC⟩.

When creating new candidate utterances, every identified entity is filled into each template of the same type. To maintain scalability and performance, we limit encoding for up to 2K utterances to produce a response in a reasonable time.

D. Context / Response Encoder and Ranking

We leverage the open-sourced code from [22], which is a retrieval system trained on the Image-Chat dataset, and modify it by incorporating the emotional context with retraining. The architecture ranks candidate responses against the input context via dot product. BLIP [27] is utilized to encode the image and the conversation history, each utterance combined with 3-dimensional VAD emotional state, and a 1-hot encoder with a linear layer for the desired output style. The information is forwarded to a summation multimodal combiner to obtain the input context. We also tested the multimodal attention combiner, but the performance was not satisfying.

The candidate encoder employs a transformer with parameters similar to those of the context encoder but

with different weights. Candidates are obtained from a static preprocessed data bank obtained from the positive utterances of Image-Chat's training set and a dynamic set incoming from the generated templates using the recognized entities.

The overall score (s) for the response candidates (C) can be computed as:

$$s(I, S, D, C) = (r_I + r_S + r_D) \cdot r_C \tag{1}$$

where r_I, r_S, r_D, and r_C represent the encoded features for the image, style, dialogue history, and candidates, respectively. However, after offline testing the system and considering the reminiscence scope, we conclude that slightly giving more importance to dialogue than the image yields more engaging conversations. Thus, the multimodal combiner uses the following formula:

$$s(I, S, D, C) = (0.9r_I + r_S + 1.1r_D) \cdot r_C \tag{2}$$

E. Candidate Response Filter

To improve the verbal interaction between the robot and the user, we ignore utterances that are over K words long, where our experiments set K=20. Longer sentences incoming from public datasets that can be acceptable for a textual environment are not suited for our task. We aim to let people expand on their life stories, not having the robot lecture them.

Since different inputs can lead to the same output, we also remove utterances (and their templates if applicable) that have already been spoken by the robot. This avoids repetition and forces more variability in the dialogue.

F. Entity Correction

Retrieval systems face the challenge of rigidity in their utterances. Some of the selected sentences can contain an appropriate context but wrong entities. For example, when saying that I went on a two-day hiking trip, the robot can respond with: "Day 5 of our journey, today's task: hiking and climbing". The hiking context is appropriate, but there is a mismatch in the number of days. Correcting these situations allows for a more vivid interaction.

The entity correction module uses spaCy [23] to analyze the selected output utterance and replace its entities with those identified in the user's last three utterances. We utilize spaCy's 18 entity types for this process. If multiple entities of the same type are detected, the most recent one is selected, or, if mentioned in the same sentence, one is chosen arbitrarily.

V. EXPERIMENTS AND RESULTS

A. Experimental setup

The photo-reminiscence system is meant to be used on-site with the user interacting with a Zenbo Junior robot and an Android device. Employing Google's speech-to-text API and Zenbo's integrated text-to-speech, the system can successfully interact with English and Mandarin speakers. A computer with RTX 2080 Ti GPU serves as the server to run the required neural networks.

A total of 42 individuals aged from 22 to 84 years old and proceeding from four countries (US, Singapore, Taiwan, and Mexico) participated in the trials. The system language is set to English for young participants from overseas and to Mandarin for the local elders. Specifically, we conduct experiments with a total of 15 participants, comprising 5 males and 10 females, with an average age of 72.33 years (range: 51-84 years). The distribution of cognitive impairment among participants includes 2 individuals with Alzheimer's disease (AD), 5 with mild cognitive impairment (MCI), and 8 in the healthy control group (HC). Each participant was required to provide 3 to 5 photos of their liking, with the instruction of being related to happy memories. The image understanding module automatically tags the event and place of the image, which shapes the selected questions. However, the tags are manually modified to reduce noise in the pipeline and focus on the performance of the response generation system.

The gathering of experimental data from human subjects has been sanctioned by the NTUH Research Ethics Committee D, under ethics board protocol number 202207011RIND, as of August 5, 2022.

B. Conversation Flow

To mimic a reminiscence-based conversation between two humans, the robot starts the interaction by asking a question related to the content of the picture. The person is expected to answer it, to which the robot will provide a follow-up response. If the person replies again, a new response is generated. This process happens up to three times before asking a new question, bringing the topic back to the content of the image. However, if the person remains silent for a few seconds after a robot's response, it indicates there's nothing else to say, and the system asks the next question to keep the conversation flowing.

The robot can ask between 7-10 questions before suggesting the user change the picture, which they can accept or reject. To further enhance the friendly experience, the photo-reminiscence app also supports being interrupted while talking or repeating the last utterance upon request.

C. Tested systems

We take our lab's previous work presented in [20] as the baseline, which focuses mostly on asking questions. Adding the response generation module should lead to more engaging and pleasant interactions with people.

During the offline examination, it was noticeable that the system might struggle to provide an appealing response when the user responds with a simple "yes" or "no" due to the lack of context. To handle this case, we employ a heuristic strategy in which the robot will continue the conversation by choosing the next question instead of generating a response. However, for a fluent

TABLE I

Perspectives to evaluate the response systems

Perspective	Description	Utilized Statements
Engagingness	How engaging the user feels the system.	• The system is engaging • The responses made by the system are interesting and captivating. • I would enjoy talking with the system for a long time.
Relatedness	The responses are related to the image and user's replies.	• The generated responses are related to each other. • The generated responses are relevant to the image. • The system understands my replies.
Effectiveness	The system helps the user reminiscence about their past experiences.	• Chatting with the system helps me recall memories related to the photo. • Chatting with the system helps me remember past experiences better. • Chatting with the system reminds me of happy memories.
Appropriateness	The responses are suitable for the conversation.	• The generated responses are appropriate to the conversation. • The responses seem on-topic with the conversation. • The system is responding adequately to what I say.

TABLE II

Questionnaire ratings per perspective

	Engagingness		Relatedness		Effectiveness		Appropriateness	
System	Average	Variance	Average	Variance	Average	Variance	Average	Variance
Work of [22]	3.45	1.24	3.45	1.21	3.85	1.30	3.52	1.10
Work of [20]	3.56	0.61	3.89	0.93	4.11	0.58	3.94	0.76
Ours	4.33	0.66	4.18	0.94	4.31	0.66	4.28	0.89

TABLE III

Paired t-test between "Work of [20]" and "Ours"

	Engagingness	Relatedness	Effectiveness	Appropriateness
t Stat	3.010	1.317	0.825	1.844
t Critical two-tail	2.110	2.110	2.110	2.110
P(T≤t) two-tail	0.008	0.205	0.421	0.083

chat flow, users are requested to provide more complete responses over brief replies.

Participants interact with each system for at least 5 minutes separately, and the examinations are carried out in the following order:

- "Work of [22]": The best retrieval model proposed by Shuster et al, which expects as input an image, the conversation history, and the preferred style for the output response.
- "Work of [20]": The lab's previous reminiscence system mainly focusing on questions;
- "Ours": The presented photo-reminiscence system with automated response generation.

D. Evaluation and Results

Using a Google Forms questionnaire, participants rate the system on a Likert-like scale ranging from 1 (worst) to 5 (best), aiming at evaluating four different perspectives, each with three questions, as in Table I. These are described as: 'Engagingness', how engaging the user feels the system; 'Relatedness', the responses are related to the image and user's replies; 'Effectiveness', the system helps the user reminiscence about their past experiences; and 'Appropriateness', the responses are suitable for the conversation.

With 42 participants and 3 questions for each perspective, we get a total of 126 data points per perspective. Results are displayed in Table II. The ratings from the proposed system are higher than the work of [20] across all metrics, indicating the usefulness of the response generation module. Besides, when reviewing the dialogue records, it is found that having additional context in the input sequence allows the system to gather extra information and provide more interesting, pleasing, or captivating responses.

Statistical significance was also checked using two-tailed paired t-tests. The test returns a t Stat, which is then compared against a two-tail t Critical value. If the absolute value of t Stat is larger than t Critical, we can reject the null hypothesis and say the results are significant. To simplify the process, the P value, listed as "P(T≤t) two-tail", is considered significant if it is lower than a threshold. Table III displays the comparison performed between "Work of [20]" and "Ours", as they are the systems of most interest. Engagingness is the only perspective that is statistically significant when setting the threshold to 0.05, i.e., $p<0.05$. This result highlights the importance of providing responses for a natural, captivating, and appealing conversation. Despite both systems being highly effective at triggering reminiscence (as seen in Table II), a methodology that mainly asks questions can fail to mesmerize or promote dialogue and become boring over time. The proposed response strat-

egy increases the usefulness of the system, especially in the long run. Even though the rest of the metrics are not statistically significant, the systems seem well-received by the public. Scores are higher than the expected average, indicating they can entail communication and effectively trigger reminiscence to different degrees.

VI. CONCLUSIONS

This work introduces an empathetic response generation system for a photo-reminiscence chatbot, aiming for natural and engaging conversations. By integrating a multimodal transformer-based retrieval chatbot with a template generation mechanism, the system provides relevant follow-ups based on mentioned names or places. An emotion analysis module tailors responses according to the user's emotional state and image context. Participants, interacting through a natural speech interface, rated the system's performance on a Likert-like scale, achieving higher engagingness, relatedness, effectiveness, and appropriateness scores compared to a question-focused system, with statistical significance in engagement. Future work will involve testing with a larger group of cognitively impaired elders and expanding dialogue capabilities using a knowledge base of user information.

References

[1] W. H. O. WHO, "Dementia," 21 September 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed 1 August 2021].

[2] L. O' Philbin, B. Woods, E. Farrell, A. Spector and M. Orrell, "Reminiscence therapy for dementia: an abridged Cochrane systematic review of the evidence from randomized controlled trials," Expert Review of Neurotherapeutics, vol. 18, 2018.

[3] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster and J. Urbanek, "The Second Conversational Intelligence Challenge (ConvAI2)," The NeurIPS '18 Competition, pp. 187-208, 2020.

[4] J.-J. Wang, M. Yen and W.-C. OuYang, "Group reminiscence intervention in Taiwanese elders with dementia," Archives of gerontology and geriatrics, no. 49, 2009.

[5] K.-J. Chiang, ,. R.-B. Lu, H. Chu, Y.-C. Chang and K.-R. Chou, "Evaluation of the effect of a life review group program on self-esteem and life satisfaction in the elderly," International journal of geriatric psychiatry, Vols. 23,1, pp. 7-10, 2008.

[6] D. Cruz-Sandoval and J. Favela, "A Conversational Robot to Conduct Therapeutic Interventions for Dementia," IEEE Pervasive Computing, vol. 18, pp. 10-19, 2019.

[7] D. Cruz-Sandoval, A. Morales-Tellez, E. B. Sandoval and J. Favela, "A Social Robot as Therapy Facilitator in Interventions to Deal with Dementia-related Behavioral Symptoms," HRI '20: In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 161-169, March 2020.

[8] A. A. Ryan, C. O. McCauley, E. A. Laird, A. Gibson, M. D. Mulvenna, R. Bond, B. Bunting, K. Curran and F. Ferry, ""There is still so much inside': The impact of personalised reminiscence, facilitated by a tablet device, on people living with mild to moderate dementia and their family carers," Dementia (London, England), Vols. 19,4, pp. 1131-1150, 2020.

[9] C. O. McCauley, R. B. Bond, A. Ryan, M. D. Mulvenna, L. Laird, A. Gibson, B. Bunting, F. Ferry and K. Curran, "Evaluating User Engagement with a Reminiscence App Using Cross-Comparative Analysis of User Event Logs and Qualitative Data," Cyberpsychology, behavior and social networking, Vols. 22,8, pp. 543-551, 2019.

[10] L. Asprino, A. Gangemi, A. G. Nuzzolese, V. Presutti and A. Russo, "Knowledge-driven Support for Reminiscence on Companion Robots," AnSWeR@ESWC, Application of Semantic Web technologies in Robotics, pp. 51-55, 2017.

[11] L. Asprino, A. Gangemi, A. Giovanni Nuzzolese, V. Presutti, D. Reforgiato Recupero and A. Russo, "Ontology-Based Knowledge Management for Comprehensive Geriatric Assessment and Reminiscence Therapy on Social Robots," Data Science for Healthcare: Methodologies and Applications, pp. 173-193, 2019.

[12] M. Carós, M. Garolera, P. Radeva and X. Giro-i-Nieto, "Automatic Reminiscence Therapy for Dementia," ICMR '20: In Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 383-387, June 2020.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[14] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," In Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 3104-3112, 2014.

[15] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He and L. Vanderwende, "Generating Natural Questions About an Image," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1802-1813, 2016.

[16] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 2204-2213, 2018.

[17] Y.-L. Wu, E. Gamborino and L.-C. Fu, "Interactive Question-Posing System for Robot-Assisted Reminiscence From Personal Photographs," IEEE Transactions on Cognitive and Developmental Systems, vol. 12, no. 3, pp. 439-450, 2020.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations (ICLR), 2015.

[19] K. Ahmad, N. Conci and G. Boato, "USED: a large-scale social event detection dataset," The 7th International Conference on Multimedia and Expo, pp. 1-6, 2016.

[20] E. Gamborino, A. Herrera, J.-F. Wang, T.-Y. Tseng, S.-L. Yeh and L.-C. Fu, "Towards Effective Robot-Assisted Photo Reminiscence: Personalizing Interactions Through Visual Understanding and Inferring," HCII 2021: Cross-Cultural Design. Applications in Cultural Heritage, Tourism, Autonomous Vehicles, and Intelligent Agents, pp. 335-349, 2021.

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 1452-1464, 2018.

[22] K. Shuster, S. Humeau, A. Bordes and J. Weston, "Image-Chat: Engaging Grounded Conversations," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 2414–2429, July 2020.

[23] M. Honnibal and I. Montani, "spaCy meets Transformers: Fine-tune BERT, XLNet and GPT-2," Explosion AI, 2 Aug 2019. [Online]. Available: https://explosion.ai/blog/spacy-transformers. [Accessed 18 June 2021].

[24] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh, "Dimensional Emotion Detection from Categorical Emotion," In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4367-4380, Nov 2021.

[25] H.-Y. Chang, "Self-supervised Guided Modality Disentangled Representation Learning for Multimodal Sentiment Analysis and Schizophrenia Assessment," M.S. thesis, National Taiwan University (NTU), 2023.

[26] E. M. Smith, D. Gonzalez-Rico, E. Dinan, and Y.-L. Boureau, "Controlling Style in Generated Dialogue," ArXiv, 2020.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," In International conference on machine learning, pp. 12888-12900, PMLR, June 2022.