語音服務自主定位、導覽與情緒辨識擬真人物互動機器人

指導教授：李國川教授

學生：盧易聖、馮守誠、呂育傑

國立聯合大學 資訊工程學系

苗栗市南勢里聯大 2 號

gclee@nuu.edu.tw

摘要

本研究提出一套搭載於輪式機器人平台的嵌入式情感陪伴系統，深度整合了臉部情緒辨識、情感對話微調模型，以及擬真語音與虛擬角色生成技術，旨在實現具備高度沉浸感的人機互動。

在視覺感知方面，系統以 Vision Transformer (ViT) 為主幹網路，並結合動作單元 (AUs) 偵測機制，有效提升對細微表情的捕捉能力，於 RAF-DB 資料集上達到 92% 的辨識準確率。在認知與對話生成層面，採用 MediaTek/Breeze-7b 作為核心語言模型，利用 Empathetic Dialogues 資料集進行微調，並導入檢索增強生成 (RAG) 架構與 WordNet 情緒映射機制，使機器人能生成精確且具同理心的中文回應。為進一步強化互動真實感，系統透過 GPT-SoVITS 實現少樣本情感語音合成，並結合 EchoMimic 生成與語音口型同步的虛擬角色影片。實驗評估結果顯示，微調後的模型在 BERT F1 與 Cosine Similarity 指標上分別提升了 26.35% 與 85.67%，證實本系統能有效具備情境感知能力，並提供流暢且富含情感的陪伴體驗。

關鍵字：生成式人工智慧、深度學習、智慧機器人

Abstract

This study proposes an embedded affective companion system deployed on a wheeled robot platform, deeply integrating facial emotion recognition, fine-tuned emotional dialogue models, and realistic speech and virtual avatar generation technologies to achieve highly immersive Human-Robot Interaction (HRI).

Regarding visual perception, the system utilizes a Vision Transformer (ViT) as the backbone network combined with an Action Unit (AU) detection mechanism. This approach effectively enhances the capture of subtle facial expressions, achieving a recognition accuracy of 92% on the RAF-DB dataset. For cognition and dialogue generation, MediaTek/Breeze-7b is adopted as the core language model. By fine-tuning on the Empathetic Dialogues dataset and incorporating a Retrieval-Augmented Generation (RAG) architecture alongside a WordNet-based emotion mapping mechanism, the system enables the robot to generate precise and empathetic Chinese responses. To further enhance interaction realism, the system implements few-shot emotional speech synthesis via GPT-SoVITS and utilizes EchoMimic to generate virtual avatar videos with precise lip-synchronization. Experimental results demonstrate that the fine-tuned model achieved improvements of 26.35% in BERT F1 and 85.67% in Cosine Similarity. These findings confirm the system's effective contextual awareness and its ability to provide a fluent, emotionally rich companion experience.

Keyword：Generative AI, Deep Learning, Intelligent Robots

一、研究動機及目的

隨著人工智慧與自然語言處理技術的快速演進，生成式 AI 在心理支持與社交陪伴領域展現了巨大潛力，然而現有的互動系統多偏重於功能性的任務輔助，往往缺乏即時的情緒感知能力與具備溫度的情感表達，導致人機互動難以建立深層的連結與信任。傳統的社交機器人受限於感知演算法在細微表情辨識上的不足，以及通用語言模型在同理心回應上的侷限，常使互動過程顯得生硬且缺乏沉浸感，無法滿足使用者在情感慰藉上的真實需求。

基於上述挑戰，本研究旨在開發一套深度整合多模態感知與生成技術的嵌入式情感陪伴機器人系統，以突破現有技術在情感理解與表達上的瓶頸。透過結合 Vision Transformer 與動作單元偵測技術以提升情緒辨識的精確度，並利用情感對話資料集微調大型語言模型以強化對話的同理心邏輯，同時整合擬真語音合成與虛擬角色生成技術，建構出涵蓋視覺、聽覺與語意理解的完整互動迴圈。本研究期望藉由搭載於具自主導航能力的實體機器人平台，實現能主動感知使用者情緒並提供適時心理支持的智慧陪伴應用，為未來的人機情感互動發展奠定堅實基礎。

二、系統架構及流程

1. 系統架構

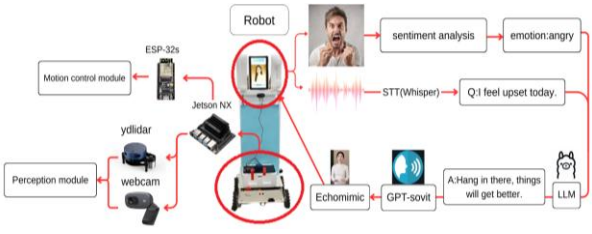

圖 1.系統架構圖

本系統採用三層式架構如圖 1 所示。邊緣端由 Jetson NX 負責，主要執行 YOLO 演算法進行人臉定位，並透過 LiDAR 與 Hector-Mapping SLAM 技術完成環境建圖與導航[20]，以支援機器人的自主移動與感知。前端互動層則作為使用者與系統之間的介面，接收使用者輸入，並輸出對應的虛擬角色影像回應，實現自然的人機互動。後端伺服器則承擔運算需求較高的工作，包括微調後的大型語言模型、GPT-SoVITS 語音合成，以及 EchoMimic 虛擬角色生成。三層架構透過網路與 API 進行連結與協作，共同完成即時感知、情緒辨識、多模態生成與人機互動等任務。

三、個別功能介紹

1. 臉部情緒辨識

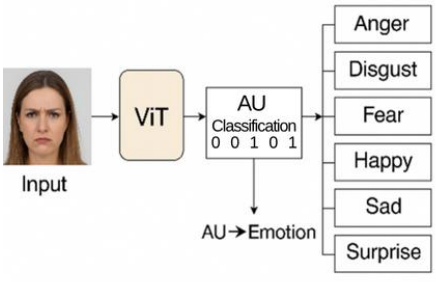

圖 2. 情緒辨識模組整體流程

本研究之情緒辨識模組（流程如圖 2 所示）採用 Vision Transformer (ViT) [1] 作為特徵提取主幹，透過將人臉影像區塊化並結合自注意力機制，有效突破傳統 CNN 之侷限，能同時捕捉全域結構語意與局部細微特徵。在分類機制上，本系統不直接預測情緒，而是引入基於 FACS [16] 的動作單元 (Action Units, AUs) 作為中介特徵；將 AU 偵測建模為多標籤分類任務，並利用 Sigmoid 激活函數與二元交叉熵損失進行優化。此設計使模型能精確捕捉眉眼或嘴角等局部肌肉的微小變化，最後再將 AU 組合映射至七類基本情緒，顯著提升了系統在跨樣本情境下的辨識穩健性與泛化能力 [27]。

2. 語言模型與情感對話生成

本模組以 MediaTek/Breeze-7b [5] 為核心，利用 Empathetic Dialogues [3] 進行同理心微調，並整合檢索增強生成 (RAG) 架構 [4] 以引入專業心理輔導知識，解決通用模型在特定領域知識的不足。透過此機制，系統能生成兼具情境感知、正確性與情感支持的回應，並作為後續情感語音

合成的輸入基礎，實際運作流程如圖 3 所示。

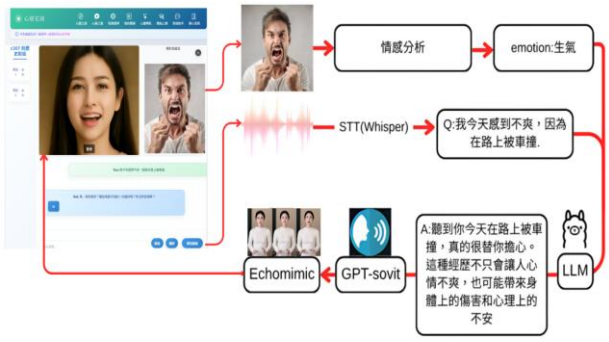

圖 3. 語言模型與情感對話生成

3. 臉部情緒與對話情緒的映射

為解決對話資料集之細粒度情緒標籤與視覺模組通用分類間的不一致，本研究提出一種基於 LLM 的自動化語意映射機制。有別於傳統人工規則，本方法整合 WordNet [31] 語意網與 Plutchik 情緒輪 [32] 理論，使模型能依據同義詞集與情緒維度進行邏輯推論。系統能自動將高喚醒度與攻擊性詞彙（如 furious, jealous）映射至「憤怒」;將具排斥與自我否定意涵者(如 guilty, disappointed）歸類為「厭惡」; 並將正向效價或低能量詞彙分別導向「快樂」或「悲傷」等六大基本類別（完整對照如圖 5)。此機制不僅實現了文本到視覺表情的精準轉換，更確保系統在面對未見情緒詞彙時具備良好的泛化理解能力。

圖 5 用戶照片上傳（圖像為 ChatGPT 生成）

4. 情感語音合成

為提升互動臨場感，本研究採用 GPT-SoVITS [6]進行情感語音合成，該模型具備跨語言與少樣本聲音克隆優勢。系統架構首先以 HeBERT [17] 提取音色特徵，經由 VALL-E [18] 進行語意建模與融合，最終透過 VITS 解碼器生成具備情緒表現的語音（流程如圖 4)。在微調前處理方面，音訊樣本經切分（3–10 秒）與 UVR 降噪後，利用 Whisper [19]自動標註文本對齊資料，確保模型能精準重現目標話者的聲紋與情感特徵。

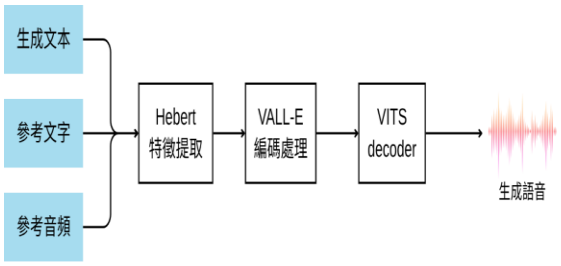

圖 4. GPT-SoVITS 流程圖

5.虛擬角色生成

虛擬角色生成模組利用 EchoMimic [7] 將 GPT-SoVITS 生成之情感語音與使用者照片進行整合（輸入如圖 6、7)，驅動具備精確口型同步與自然頭部動態的虛擬角色。為解決生成過程中的畫質損耗，系統引入 PMRF 演算法 [13] 進行影像修復與細節增強，消除模糊與噪點。整體流程被設計為全自動化管線，能在無人工干預下即時輸出高擬真動態影片（如圖 8)，顯著提升人機互動的視覺真實感與沉浸體驗。

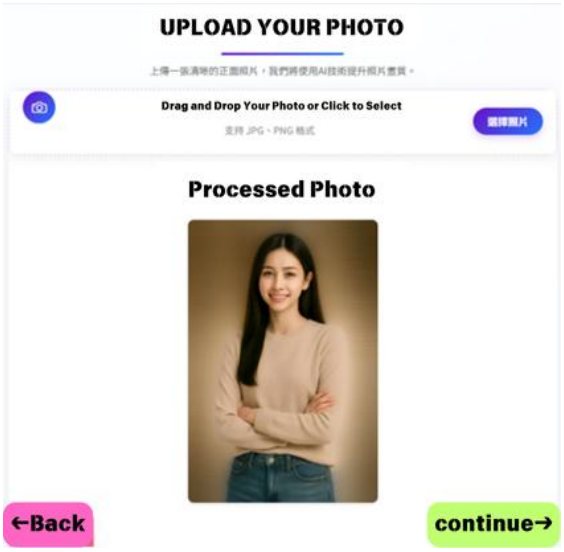

圖 6 用戶照片上傳（圖像為 ChatGPT 生成）

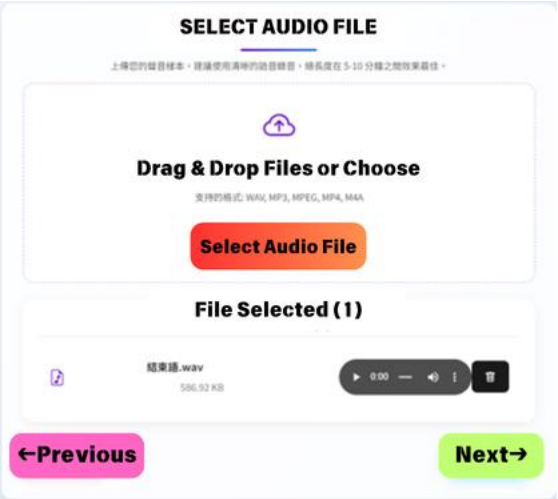

圖 7 使用者語音上傳

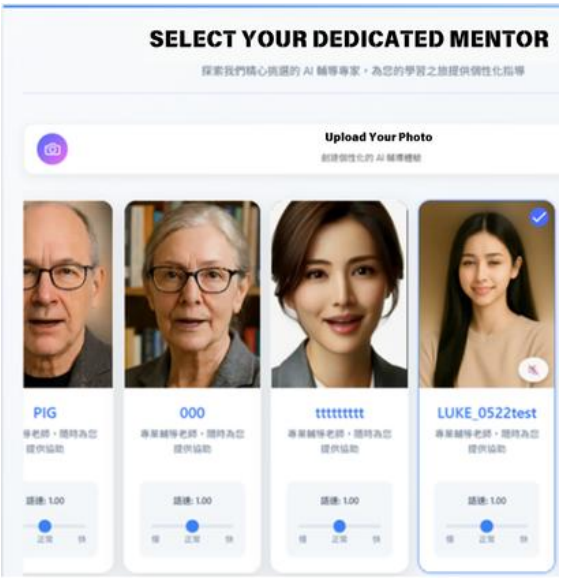

圖 8 動態虛擬角色(人像皆由 ChatGPT 生成)

6. 前端頁面設計與資料庫建立

為建構高效人機互動介面，本系統採用 Next.js 框架，利用其伺服器端渲染特性確保操作流暢性。後端資料庫採關聯式設計（如圖 9），以 users 表為核心連結 chats 表管理對話情境；messages 表除儲存文本與發送者資訊外，更特設 audio_uri 欄位索引情感語音，實現多模態數據的完整保存。此外，系統結合 chat_logs 記錄操作歷程，有效保障數據的一致性與可追溯性。

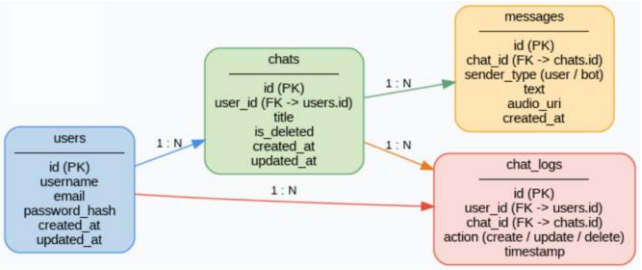

圖 9 使用者語音上傳

7. 環境感知與即時地圖生成

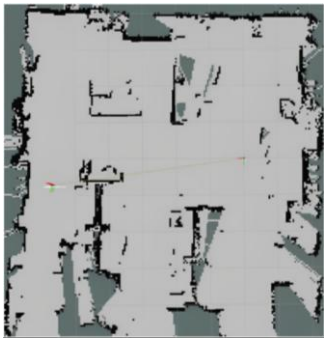

圖 10. 實際建立實驗室地圖

為實現未知室內環境之自主導航，本系統採用 Hector SLAM [20] 作為核心。該演算法不依賴里程計，而是透過高頻雷射掃描匹配技術，直接將即時 LiDAR 點雲與柵格地圖進行對齊。此機制能有效規避輪子打滑所致之累積誤差，確保定位精確度，生成地圖如圖 10 所示。

8. 用戶掃描與距離偵測

本系統採用輕量化 YOLO-Face [28] 進行即時人臉偵測，確保在複雜光照下的強健性，並串接 FaceNet [29] 將感興趣區域 (ROI) 轉換為 128 維特徵向量。透過計算與預存特徵之歐式距離進行身分驗證，僅在符合閾值時觸發互動。確認身分後，系統利用雙目視差 [30] 或感測融合技術解算使用者之距離與方位（如圖 11），並透過座標轉換矩陣將其映射至全域 SLAM 地圖作為動態導航目標。此流程整合了視覺辨識與空間定位，實現了具備身分感知的高精度引導與互動。

圖 11 雙目視差計算圖

9. 視覺感知與定位

為賦予機器人精確的三維空間感知，本研究採用雙目立體視覺技術，並針對鏡頭畸變與光軸偏差問題，參考 Batpurev [34] 與張正友標定法 [35] 進行系統校正。實驗採用內角點（方格邊長 3.19 cm）之棋盤格作為參照（如圖 12），透過 OpenCV 演算法庫解算內參與畸變係數，並利用 stereoRectify 進行極線校正，將立體匹配簡化為一維水平搜尋。雖理論深度 Z 可由公式 $Z=\frac{f*B}{d}$ 求得（其中 f 為焦距，B 為基線，d 為視差），但考量硬體組裝導致的系統性誤差，本研究進一步對 60cm 至 720cm 之測距範圍進行線性回歸分析，推導出修正後的深度公式 $Z_{linear}=\alpha*Z+ß$。此一整合標定、校正與回歸補償的流程，有效消除了非線性光學誤差，為後續導航定位奠定高可靠度的三維座標基礎。

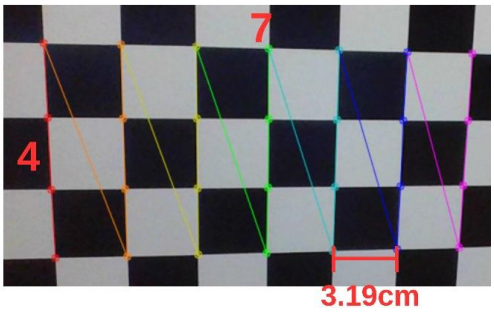

圖 12 棋盤格特徵點偵測示意圖

11. 專題實驗

本研究以 RAF-DB 與繁體中文版 EmpatheticDialogues 資料集分別作為視覺情緒辨識與情感對話生成的實驗基準。在視覺模組方面，採用 Vision Transformer (ViT) 為主幹網路並結合動作單元 (AU) 偵測機制，以強化對細微表情特徵的捕捉；語言模型則基於 MediaTek/Breeze-7b，利用 LoRA 技術進行指令式微調，使模型能建立語境與情緒標籤的深層關聯。此外，針對實體機器人的導航需求，本研究亦在 NVIDIA Jetson NX 嵌入式平台上，對比了 A* 與 Dijkstra 演算法在路徑規劃上的運算效能。

實驗結果顯示，本研究提出的 ViT 結合 AU 架構在情緒辨識準確率上達到 92.0%，顯著優於 ResNet-50 與純 ViT 模型，證實了動作單元特徵能有效提升辨識效能。在對話生成方面，微調後的模型在 BERT F1 與 Cosine Similarity 指標上分別提升了 26.35% 與 85.67%，量化數據充分證明了針對特定情緒語料進行微調，能大幅增強語言模型的情感感知能力，使其生成更具同理心與語意契合度的回應。

四、結論

本研究旨在突破傳統人機互動系統在情感深度與實體感知上的侷限，成功開發了一套整合感知、認知、表達與行動能力的具身化情感陪伴機器人。在感知與認知層面，實驗證實採用 Vision Transformer 結合動作單元偵測的階層式架構，能顯著提升複雜情緒的辨識準確率；配合經 WordNet 情緒映射與同理心語料微調的 Breeze-7B 語言模型，有效解決了通用模型回應僵化的問題，賦予系統具備高度同理心的中文對話能力。

在多模態表達與實體互動方面，本系統透過 GPT-SoVITS 與 EchoMimic 技術實現了具備個人化聲音克隆與高擬真口型同步的虛擬角色，並結合 PMRF 影像修復技術大幅增強互動沉浸感。同時，基於 ROS 架構整合 Hector SLAM 與 A* 路徑規劃演算法，賦予機器人在複雜環境中自主避障與主動接近使用者的能力，實現真正的「主動陪伴」。本研究所提出的多模態整合框架，不僅驗證了在真實場域提供高品質情感支持的可行性，亦為未來智慧長照與心理支持機器人的發展提供了具體的實踐。

五、專題影片

https://youtu.be/rfVrRYHdm1E

六、參考文獻

[1] A. Dosovitskiy, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.

[2] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 356–370, 2019.

[3] H. Rashkin, E. M. Smith, M. Li, and Y. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in Proc. ACL, 2019, pp. 5370–5381.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. NeurIPS, 2020.

[5] C.-J. Hsu, C.-L. Liu, F.-T. Liao, P.-C. Hsu, Y.-C. Chen, and D.-S. Shiu, “Breeze-7B technical report,” arXiv preprint arXiv:2403.02712, 2024.

[6] RVC-Boss, “GPT-SoVITS: Cross-lingual zero-shot voice cloning,” GitHub repository, 2024. [Online]. Available: https://github.com/RVC-Boss/GPT-SoVITS

[7] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, “EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditioning,” arXiv preprint arXiv:2407.08136, 2024.

[8] R. Meng, X. Zhang, Y. Li, and C. Ma, “EchoMimicV2: Towards striking, simplified, and semi-body human animation,” arXiv preprint arXiv:2411.10061, 2024.

[9] EmoLLM Team, “EmoLLM: Reinventing mental health support with large language models,” GitHub repository, 2024. [Online]. Available: https://github.com/EmoLLM

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, and R. A. Saurous, “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions,” in Proc. ICASSP, 2018, pp. 4779–4783.

[11] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in Proc. ICML, 2021.

[12] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “Wav2Lip: Accurately lip-syncing videos to arbitrary speech,” in Proc. ACM MM, 2020, pp. 812–820.

[13] G. Ohayon, T. Michaeli, and M. Elad, “Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration,” arXiv preprint arXiv:2410.00418, 2025.

[14] Team Gemma, Gemma: Open language models. Google DeepMind, 2025. [Online]. Available: https://ai.google.dev/gemma

[15] Unsloth. (2024). Unsloth: 2-5x faster LLM finetuning & 80% less memory [Software]. GitHub. https://github.com/unslothai/unsloth

[16] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Trans. Audio, Speech, and Language Processing, 2021. (arXiv:2106.07447)

[18] C. Wang, S. Chen, Y. Wu, et al., “Neural codec language models are zero-shot text to speech,” arXiv preprint arXiv:2301.02111, 2023. (VALL-E)

[19] A. Radford, J. Kim, T. Xu, et al., “Robust speech recognition via large-scale weak supervision,” arXiv preprint arXiv:2212.04356, 2022. (Whisper)

[20] S. Kohlbrecher, O. von Stryk, J. Meyer, and U. Klingauf, “A flexible and scalable SLAM system with full 3D motion estimation,” in Proc. IEEE SSRR, 2011. (Hector SLAM)

[21] P. E. Hart, N. J. Nilsson, and B. Raphael, “A formal basis for the heuristic determination of minimum cost paths,” IEEE Trans. Systems Science and Cybernetics, vol. 4, no. 2, pp. 100–107, 1968. (A*)

[22] S. Mangrulkar, L. Tunstall, P. von Platen, et al., “PEFT: State-of-the-art parameter-efficient fine-tuning methods,” GitHub repository, 2022. (Parameter-Efficient Fine-Tuning)

[23] E. Hu, Y. Shen, P. Wallis, et al., “LoRA: Low-rank adaptation of large language models,” arXiv preprint arXiv:2106.09685, 2021.

[24] Y. Yu, L. Wu, Z. Liu, et al., “RS-LoRA: Rank stabilized low-rank adaptation for large language models,” arXiv preprint arXiv:2404.00639, 2024.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in Proc. ICLR, 2020.

[26] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. (cosine similarity)

[27] W. Yan, S. Li, C. Que, J. Pei, and W. Deng, “RAF-AU Database: In-the-wild facial expressions with subjective emotion judgement and objective AU annotations,” arXiv preprint arXiv:2008.05196, 2020.

[28] Z. Zhang et al., “YOLO-Face: A Real-Time Face Detector Based on YOLO Architecture,” IEEE Access, 2021.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. CVPR, 2015.

[30] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge Univ. Press, 2004.

[31] G. A. Miller, “WordNet: A lexical database for English,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[32] R. Plutchik, “The nature of emotions,” American Scientist, vol. 89, no. 4, pp. 344–350, 2001.

[34] T. Batpurev, “Python Stereo Camera Calibrate,” GitHub repository, 2019. [Online]. Available: https://github.com/TemugeB/python_stereo_camera_calibrate.

[35] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.