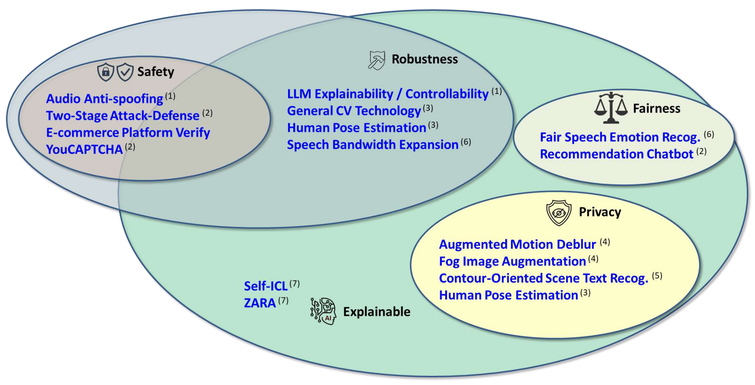
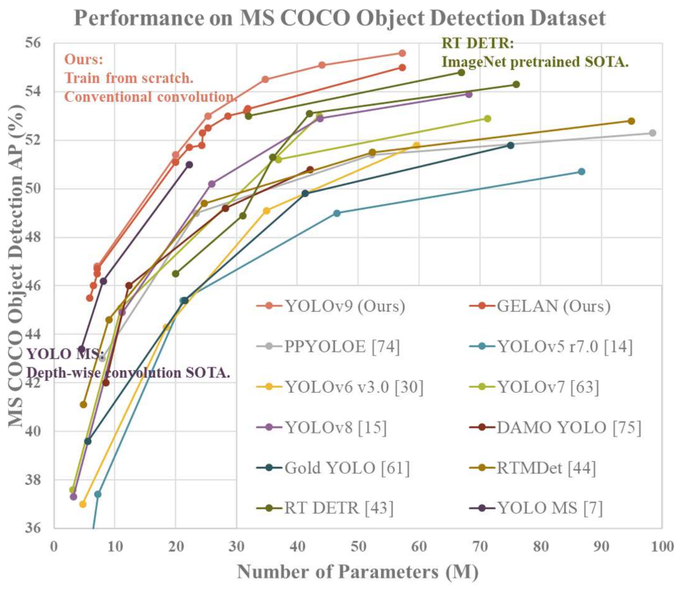
Project: Advanced Technologies for Designing Trustable AI Services
計畫名稱：適用於智慧服務的可信賴AI先進技術研究

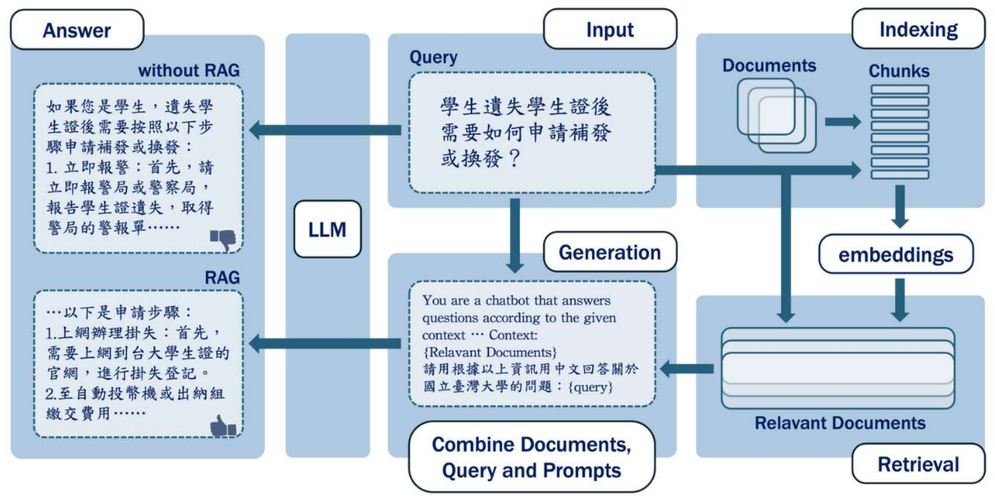
主持人/共同主持人：陳信希教授、林守德教授、吳尚鴻教授、廖弘源教授、林嘉文教授、陳祝嵩教授、李祈均教授

- Develop trustworthy advanced AI Technologies.
- Generic AI Technologies, which can be applied on various application scenarios.
- Visual content understanding
- YOLOv4, v7, v9 series release, world level impacts.





YOLOv9 performance. 2693 citations
Github 9.3k stars, 29 models



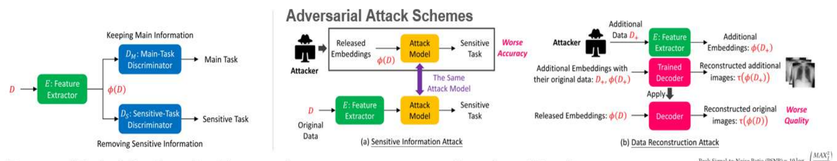
Campus Question-Answering System as a demo:
使用 GPT-4 as evaluator for accuracy
Without RAG：31%
With RAG：98%

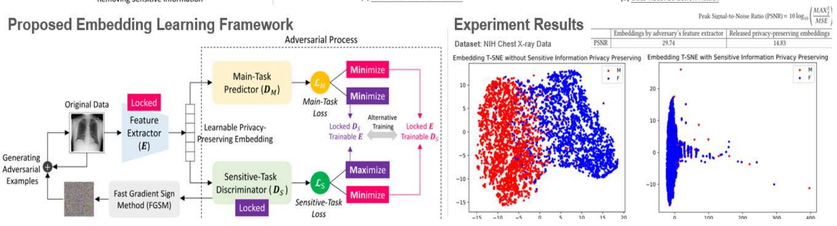
子計畫一：具可信度與可靠度的次世代AI 客服 (PI: 林守德教授，co-PI:張智星教授，陳縕儂教授)

本子計畫研究次世代AI 客服應具備幾項要素：1. 系統的隱私保護能力（左） 2. 系統解釋推論的能力 （中） 3. 對於潛在攻擊的偵測 （右）

Learning Privacy Preserving Embeddings for Image Data

Target: image D that is going to be published, along with its label → release privacy-preserving embeddings, denoted by ϕ(D).
Goal: prevent the original image data and its corresponding sensitive information hidden behind from being unveiled by the adversary.

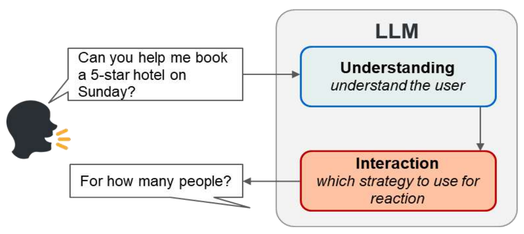




Li, Chu-Chen, Cheng-Te Li, and Shou-De Lin. "Learning Privacy-Preserving Embeddings for Image Data to Be Published." ACM Transactions on Intelligent Systems and Technology 14.6 (2023)

Injecting Expert Strategy for Explainability

Goal: 在LLM的對話互動中引入專家的互動策略 (理解使用者與決定互動方式)，開發者可理解模型為何選擇此互動方式，上升可解釋性與可控性 (Chang & Chen, ACL Findings 2024)。



A: Actually, I just finished a mystery novel that had a great plot twist. It's one of my favorite genres. Do you read mystery books?
U: Sometimes, but mostly I prefer sci-fi and fantasy. Have you read any good books in those genres?
A: (The user implicitly mentioned the intent of FindRestaurants. I should smoothly pivot the conversation to FindRestaurants topic)
Speaking of sci-fi and fantasy, have you ever been to a themed restaurant?
U: No, I haven't. Is that a thing?
A: (The user did not change the topic of FindRestaurants. I should continue the topic)
Yes! There are restaurants that are decorated to look like settings from popular sci-fi and fantasy movies and TV shows.
U: That sounds really cool! Do you know of any nearby? …

Contributions

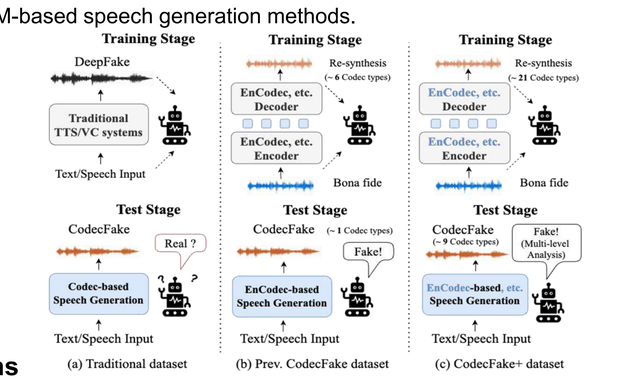
- 在多種不同使用者的模擬對話中，皆可透過此技術達成更自然也更好的行銷對話手法。
- 提出之SalesBot 2.0資料集與訓練之模型，皆開源於GitHub供後續研究使用。

Wen-Yu Chang and Yun-Nung Chen, "Injecting Salesperson's Dialogue Strategies in Large Language Models with Chain-of-Thought Reasoning," in *Findings of the Association for Computational Linguistics: ACL 2024 (Findings of ACL 2024)*, pages 3798–3812, Bangkok, Thailand, August 11-16, 2024. ACL.

Towards Generalized Audio Deepfake Detection

Background: The emerging LLM-based speech generation method poses a great threat to the traditional speech deepfake detection model.

Goal: Propose a general method for detecting and understanding audio generated by LLM-based speech generation methods.



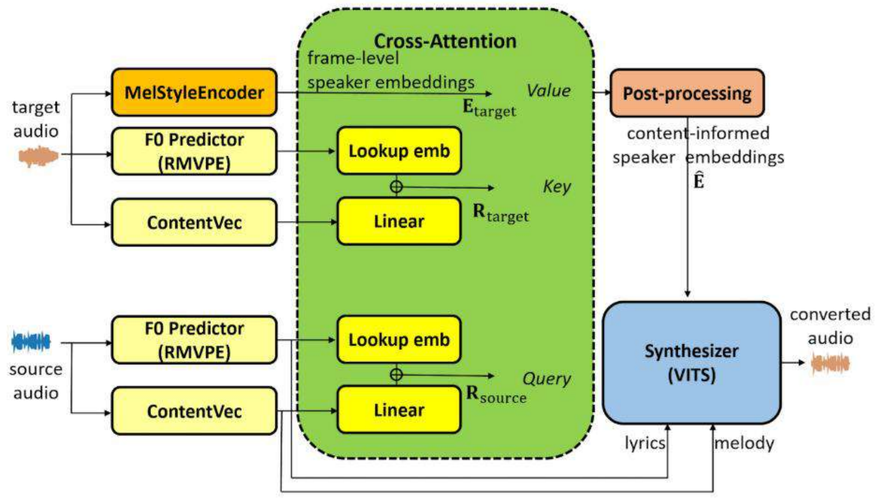
Contributions

- Propose a large-scale neural audio codec-based deepfake speech dataset
- Propose a generalized source tracing framework based on codec taxonomy

Xuanjun Chen, et al., "CodecFake+: A Large-Scale Neural Audio Codec-Based Deepfake Speech Dataset," Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.
Xuanjun Chen, et al., "Codec-Based Deepfake Source Tracing via Neural Audio Codec Taxonomy," INTERSPEECH, 2025.
Xuanjun Chen, et al., "Towards Generalized Source Tracing for Codec-Based Deepfake Speech," Submitted to ASRU 2025.

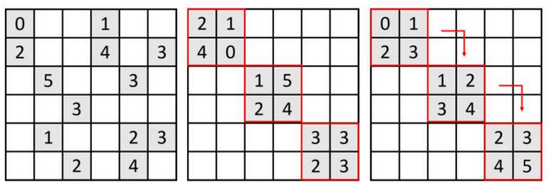
子計畫二：可信賴AI之資料處理與模型理論 (PI: 吳尚鴻教授)

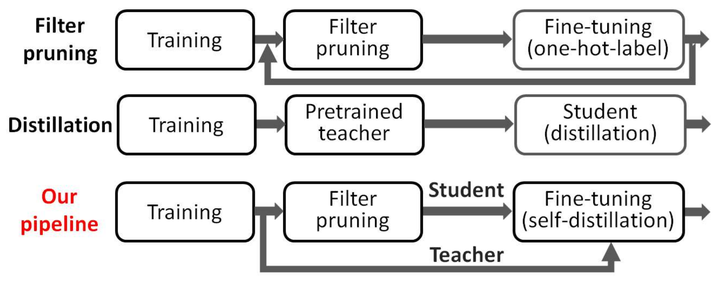
- **Yi-Wen Liu's Lab** proposes a **cross-attention-based singing voice conversion (SVC)** system that generates frame-level speaker timbre embeddings informed by content and melody. Unlike traditional SVC methods using static speaker embeddings, this approach captures dynamic vocal variations across pitch and enunciation. Results show improved clarity and voice similarity, with expressiveness remaining a future challenge.



Neural Network Compression – Shen, Chih-Ya

- **Goal:** Reduce computation and memory usage in deep neural networks.
- **FCNN Compression(Left Picture):**
  - Uses block-diagonal masking to prune weights.
  - Minimizes differences across blocks and applies delta + Huffman encoding for efficiency.
- **CNN Compression(Right Picture):**
  - Introduces "HAP self-distillation" to improve training.
  - Uses NGGM pruning (based on gradients + geometric median).
  - Employs a retraining-aware design.
- **Key Contribution:** Combines structured pruning and improved distillation methods for better compression.





子計畫三：3M Maker: 使用多模式、多任務、多事例學習開發可信賴多媒體處理技術 (PI: 廖弘源教授)

- The proposed Generalist YOLO achieves excellent results on multi-modal and multiple vision and language tasks, while maintaining higher accuracy than existing state-of-the-art generalist models with 135 times faster inference speed.
- We proposed Relaxed Optimizer and several efficient modules to design the Generalist YOLO.

Encoder	#param	FLOPs	AP^{box}	AP^{mask}	$mIoU^{sem}$	$mIoU^{stuff}$	PQ^{pano}	$B@4^{cap}$
GELAN	47.5M	~225.7G	50.3	41.8	48.7	56.8	40.4	36.8
+ Relaxed Optimizer	47.5M	~225.7G	51.9	42.6	50.0	56.5	41.7	38.8
+ PGI [35]	47.5M	~225.7G	52.1	42.6	50.1	56.4	42.1	39.1
+ Shallow Decoder	45.5M	~222.7G	52.2	42.9	49.4	56.8	42.2	**39.4**
+ Semantic Feature	37.3M	~195.2G	52.3	42.9	51.5	59.1	44.0	38.6
+ Distinctive Feature	37.3M	~195.2G	52.3	42.9	51.7	**59.4**	**44.3**	38.5
+ Light Head	**32.5M**	**~122.2G**	**52.4**	**43.0**	**52.0**	59.2	44.2	38.7

- The proposed Generalist YOLO do not need any pretrained model, and it gets prediction of all tasks in one-shot inference.

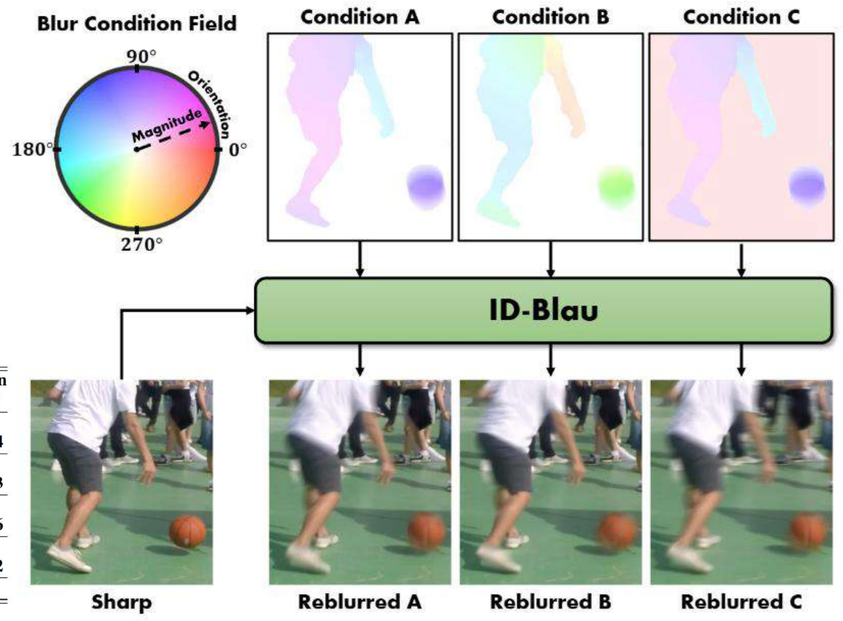
子計畫四：基於具可信賴性小樣本學習的電腦視覺技術 (PI: 林嘉文教授)

ID-Blau：運動模糊圖像的資料擴增

及新應用環境場域圖像進行 test-time adaptation 多重曝光模糊物理模型，將運動模糊進行參數化，再利用影像生成模型中的影像擴散模型(Diffusion Model)以生成更大量模糊圖像

PANet：霧景圖像生成及擴增

基於物理導引的參數擴增網路(PANet)，用少量的清晰/霧景配對圖像以學習一生成具真實分布



Model		GoPro PSNR	GoPro SSIM	HIDE PSNR	HIDE SSIM	RealBlur-J PSNR	RealBlur-J SSIM	RealBlur-R PSNR	RealBlur-R SSIM	Average Gain PSNR	Average Gain SSIM
MIMO-UNet+	Baseline	32.44	0.957	30.00	0.930	31.92	0.919	39.10	0.969		
	+ID-Blau	**32.93 (+0.49)**	**0.961**	**30.68 (+0.68)**	**0.938**	**31.96 (+0.04)**	**0.921**	**39.38 (+0.28)**	**0.971**	**+0.37**	**+0.004**
Restormer	Baseline	32.92	0.961	31.22	0.942	32.88	0.933	40.15	**0.974**		
	+ID-Blau	**33.51 (+0.59)**	**0.965**	**31.66 (+0.44)**	**0.947**	**33.11 (+0.23)**	**0.937**	**40.31 (+0.16)**	**0.974**	**+0.36**	**+0.003**
Stripformer	Baseline	33.08	0.962	31.03	0.940	32.48	0.929	39.84	0.974		
	+ID-Blau	**33.66 (+0.58)**	**0.966**	**31.50 (+0.47)**	**0.944**	**33.77 (+1.29)**	**0.940**	**41.06 (+1.22)**	**0.977**	**+0.89**	**+0.006**
FFTformer	Baseline	34.21	0.969	31.62	0.946	32.62	0.933	40.11	0.973		
	+ID-Blau	**34.36 (+0.15)**	**0.970**	**31.94 (+0.32)**	**0.949**	**32.88 (+0.26)**	**0.934**	**40.45 (+0.34)**	**0.975**	**+0.27**	**+0.002**
Average Gain		**+0.45**	**+0.003**	**+0.48**	**+0.005**	**+0.46**	**+0.005**	**+0.50**	**+0.002**	-	-

ID-Blau: Image Deblurring by Implicit Diffusion-Based reBLurring AUgmentation.
CVPR2024, ECCV2024

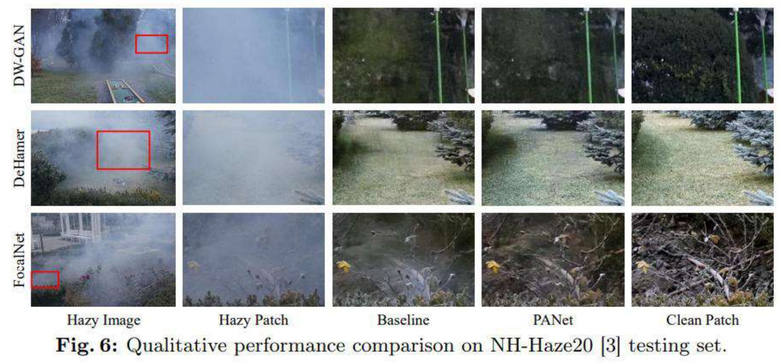


Fig. 6: Qualitative performance comparison on NH-Haze20 [3] testing set.

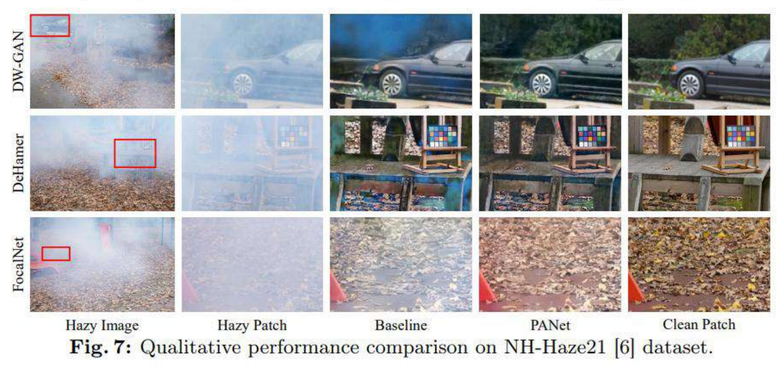


Fig. 7: Qualitative performance comparison on NH-Haze21 [6] dataset.

PANet: A Physics-guided Parametric Augmentation Net for Image Dehazing by Hazing

子計畫五：可信賴的生活場景文字識別技術 Trustworthy Text Detection and Recognition for Daily-Life Scenes (PI: 陳祝嵩教授)

Selected Publications

(CVPR'25) Relation-Rich Visual Document Generator for Visual Information Extraction. RIDGE generates relation-rich synthetic documents by aligning LLM-generated content with layout structures learned from OCR, boosting VIE performance without human annotation.

(CVPR'25) ABC-Former: Auxiliary Bimodal Cross-domain Transformer with Interactive Channel Attention for White Balance. ABC-Former fuses RGB and CIELab histograms via cross-domain attention to achieve state-of-the-art white balance correction on sRGB images.

(CVPRW'25) Syn3DTxt: Embedding 3D Cues for Scene Text Generation. Syn3DTxt introduces a new 3D-aware synthetic dataset standard using surface normals to enhance spatial realism in scene text rendering.

(ECCV'24) SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. SAM4MLLM unifies SAM and MLLMs via language-driven prompting to enable pixel-aware segmentation without architectural changes.

(EMNLP Findings'24) ACCEPT: Adaptive Codebook for Composite and Efficient Prompt Tuning. ACCEPT introduces a codebook-based prompt tuning method that achieves state-of-the-art performance with just 0.3% of PLM parameters.

Scene Text Detection & Recognition

- Scene text detection and recognition is a fundamental task in with many applications such as autonomous driving and robotic vision.
- We observed a lack of large-scale datasets for Taiwanese street scenes, which limits the effectiveness of existing OCR systems. To bridge this gap, we constructed a new annotated dataset and introduced a fine-tuned framework.

Main Achievement

- We proposed a new benchmark dataset covering **Taiwanese street and campus scenes**, addressing the lack of localized OCR datasets.



- We integrate a scene-text detection model with a MLLM for scene text recognition. This approach mitigates hallucinations commonly observed in MLLM-based OCR and achieves better performance than the Google OCR API on real-world Taiwanese street and campus scenes.

Field Validation Result

Quantitative Comparison:

TW Campus	Detection Score (IoU@0.5) P	R	F1	Recognition ANLS	End-to-end F1 * ANLS
Google OCR	0.265	0.433	0.329	0.912	0.300
Qwen2.5VL-Zero	0.639	0.429	0.514	0.782	0.402
SRFormer (FT) + Breeze2-Zero	0.638	0.647	**0.643**	0.873	**0.561**

TW Street Text	Detection Score (IoU@0.5) P	R	F1	Recognition ANLS	End-to-end F1 * ANLS
Google OCR	0.286	0.327	0.305	0.809	0.247
Qwen2.5VL-Zero	0.435	0.111	0.177	0.798	0.141
SRFormer (FT) + Breeze 2-Zero	0.596	0.618	**0.607**	0.647	**0.393**

Visualization Results:

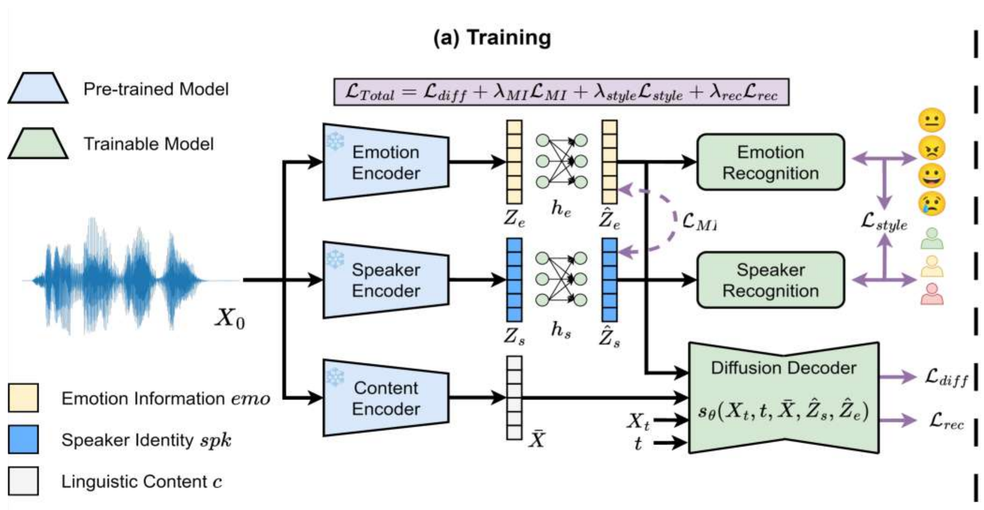




Quantitative Impact

1. We built a dataset containing 10,000+ images (540×360 resolution) with 271,842 text annotation, including: 133,459 Traditional Chinese characters, 90,523 English characters.
2. Our method's superior performance over Google OCR underscores the value of our datasets and validates the robustness of our proposed framework.
3. 15+ papers at flagship conferences (CVPR, ECCV, AAAI, ICLR, ICASSP), 30+ total publications including journals (IEEE Trans. Multimedia, Internet of Things, IEEE J. of Oceanic Engineering).

子計畫六：ZSDEVC: Zero-Shot Diffusion-based Emotional Voice Conversion with Disentangled Mechanism (PI: 李祈均教授)



Methods	MSP-Podcast UTMOS	SECS	DNSMOS SIG	DNSMOS OVRL	ECA	ESD UTMOS	SECS	DNSMOS SIG	DNSMOS OVRL	ECA
Source	2.830	1.000	3.401	2.892	0.621	3.927	1.000	3.479	3.191	0.951
Source (VO)	2.488	0.974	3.422	2.907	0.608	3.538	0.971	3.500	3.210	0.881
EMOCONV-DIFF	2.477	**0.837**	**3.528**	**3.096**	0.500	3.708	**0.822**	**3.572**	**3.315**	0.453
Our method ($\mathcal{L}_{Total}$)	2.427	0.773	3.505	3.073	0.584 (16.8 % ↑)	3.687	0.763	3.561	3.300	0.548 (21.1 % ↑)
w/ EG_{spk}	**2.493**	0.788	3.503	3.068	0.557 (11.4 % ↑)	**3.729**	0.774	3.562	3.302	0.504 (11.3 % ↑)
w/ EG_{emo}	2.353	0.744	3.485	3.040	**0.699** (40.0 % ↑)	3.665	0.747	3.560	3.298	**0.622** (37.4 % ↑)
w/ $EG_{spk,emo}$	2.383	0.766	3.484	3.032	0.672 (34.4 % ↑)	3.699	0.763	3.562	3.300	0.580 (28.1 % ↑)

Architecture: Encoders

Speech Representation Modules

- Lexical Content: Phoneme encoder (pretrained)
- Speaker Identity: Speaker encoder, speaker verification model
- Emotion Encoding: Emotion encoder, express encoder
- Input features: wav2vec with VAD (Voice Activity Detection)

Core Techniques

- Disentangled Representation: Separate speaker & emotion factors
- Guiding Mechanism: Improves expressiveness control

Diffusion-based Decoders

High quality samples

SDE

Training: predict added normal distribution noise by score model S_θ

Inference:

- Gradually remove noise
- Use sampling technique to effectively generate audios

Training Strategy: Training Objective

$L_{total} = L_{diff} + \lambda_{\mathrm{MI}} L_{MI} + \lambda_{style} L_{style} + \lambda_{rec} L_{rec}$

L_{diff}: Diffusion Loss

- MSE Loss between score function s_θ and prior noise

L_{MI}: Disentanglement Loss

- Minimize Mutual Information(MI) between speech representations

L_{style}: Style Loss

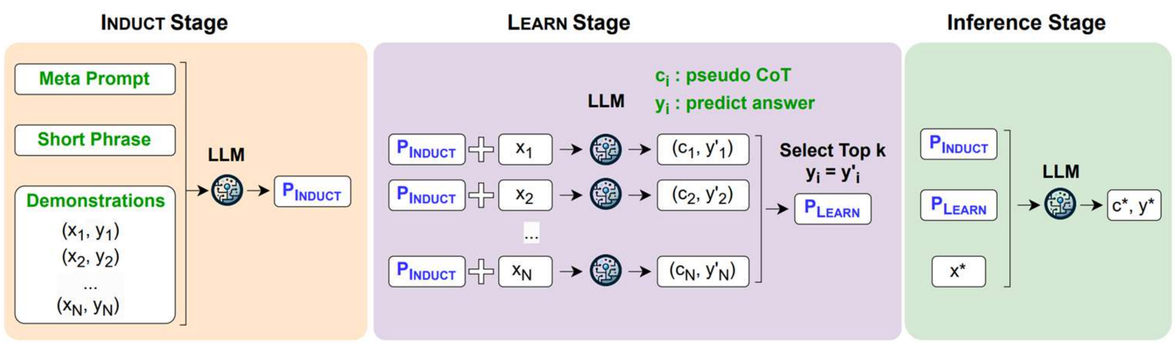
- SR/SER for style encoder

L_{rec}: Reconstruction Loss

- MSE Loss between source and estimation of source

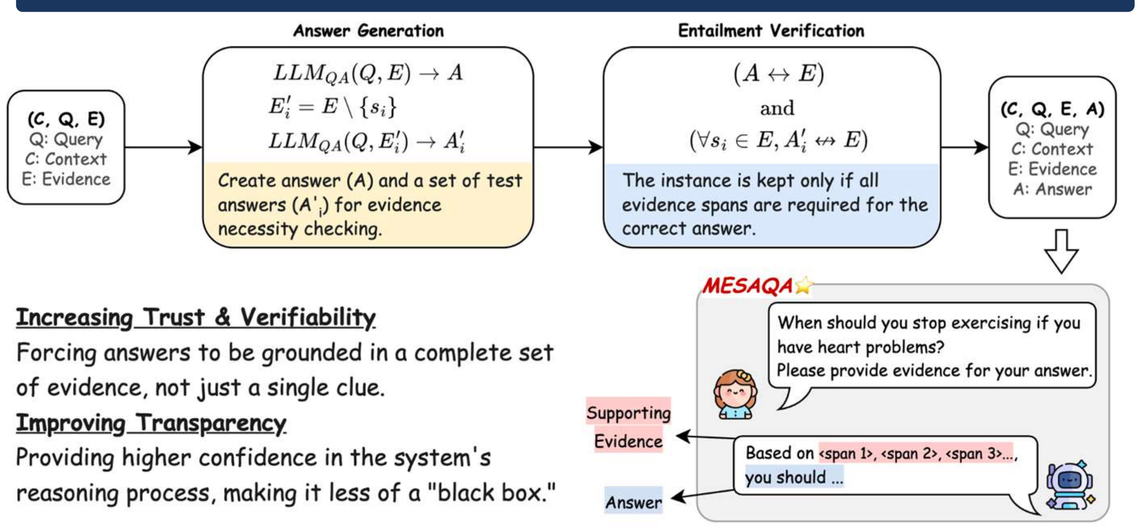
子計畫七：可信賴言談感知人類語言技術 (PI: 陳信希教授)

INDUCT-LEARN: Short Phrase Prompting with Instruction Induction



Po-Chun Chen, Sheng-Lun Wei, Hen-Hsen Huang, and Hsin-Hsi Chen
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)

MESAQA: A DATASET FOR MULTI-SPAN CONTEXTUAL AND EVIDENCE-GROUNDED QUESTION ANSWERING



Jui-I Wang, Hen-Hsen Huang and Hsin-Hsi Chen
In Proceedings of The 31st International Conference on Computational Linguistics (COLING 2025)