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

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)			Recognition	End-to-end
	P	R	F1	ANLS	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)			Recognition	End-to-end
	P	R	F1	ANLS	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).