MMM 2025

31ST INTERNATIONAL CONFERENCE ON MULTIMEDIA MODELING
JANUARY 8 - 10, 2025
NARA, JAPAN

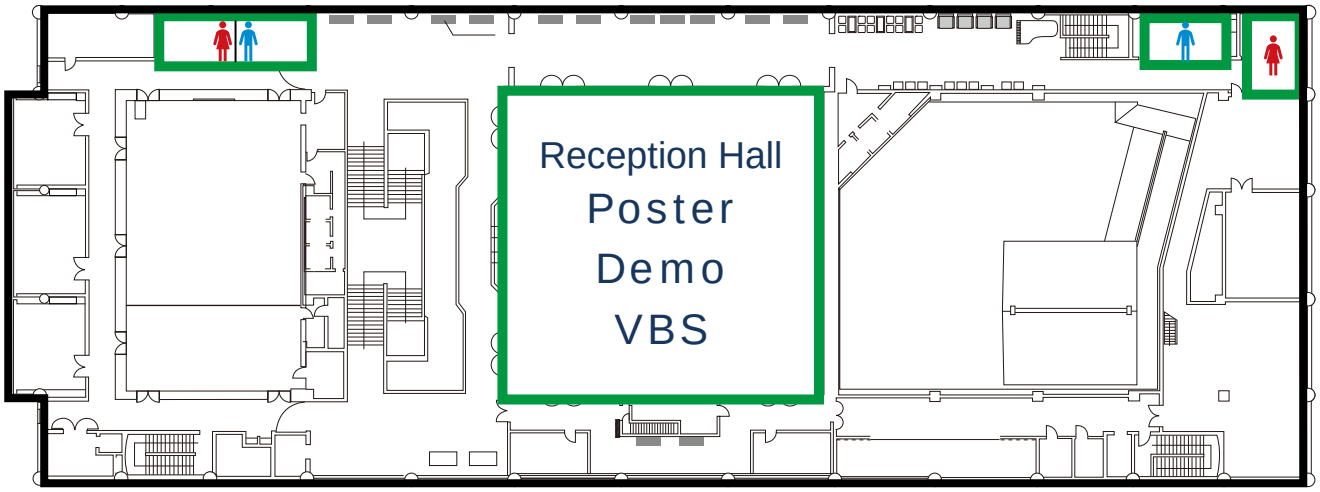
Program at a Glance

	Jan 7 (Tue)	Jan 8 (Wed)		Jan 9 (Thu)		Jan 10 (Fri)	
	2F: Reception Hall 1	1F: Noh-Gaku-Do (Noh-hall)	2F: Reception Hall 1	1F: Noh-Gaku-Do (Noh-hall)	2F: Reception Hall 1	1F: Noh-Gaku-Do (Noh-hall)	2F: Reception Hall 1
9:00		Registration @Entrance Hall		Registration @Entrance Hall		Registration @Entrance Hall	
9:15						Keynote 3 Dr. Andrei Bursuc	
9:30		Opening		Oral Session 3			
9:45		Keynote 1 Dr. Nancy F. Chen					
10:00							
10:15						Coffee Break	
10:30				Coffee Break			
10:45		Coffee Break		Oral Session 4		Oral Session 6	
11:00		Best Paper Session	Prep. for VBS (upon request)		Prep. for Demo		
11:15							
11:30							
11:45				Lunch Break		Lunch Break	
12:00		Lunch Break					
12:15							
12:30							
12:45							
13:00							
13:15			Prep. for VBS				
13:30					Poster 2 / Demo Session Coffee Break		Poster 3 / Demo Session Coffee Break
13:45							
14:00			Poster 1 Session Coffee Break				
14:15							
14:30							
14:45							
15:00				Oral Session 5		MLLMA Oral Session	
15:15							
15:30		Oral Session 1		Coffee Break			
15:45	Prep. for VBS		VBS (Expert Session)				
16:00				Keynote 2 Prof. Kiyoharu Aizawa		Coffee Break	
16:15						Oral Session 7	
16:30		Coffee Break					
16:45							
17:00		Oral Session 2				Closing	
17:15							
17:30							
17:45							
18:00			VBS (Team Presentation)	Banquet @ KOTOWA Nara Koen Premium View			
18:15							
18:30							
18:45			Welcome Reception				
19:00			VBS (Novice Session)				
19:15							
19:30							
19:45							
20:00							
20:15							
20:30							

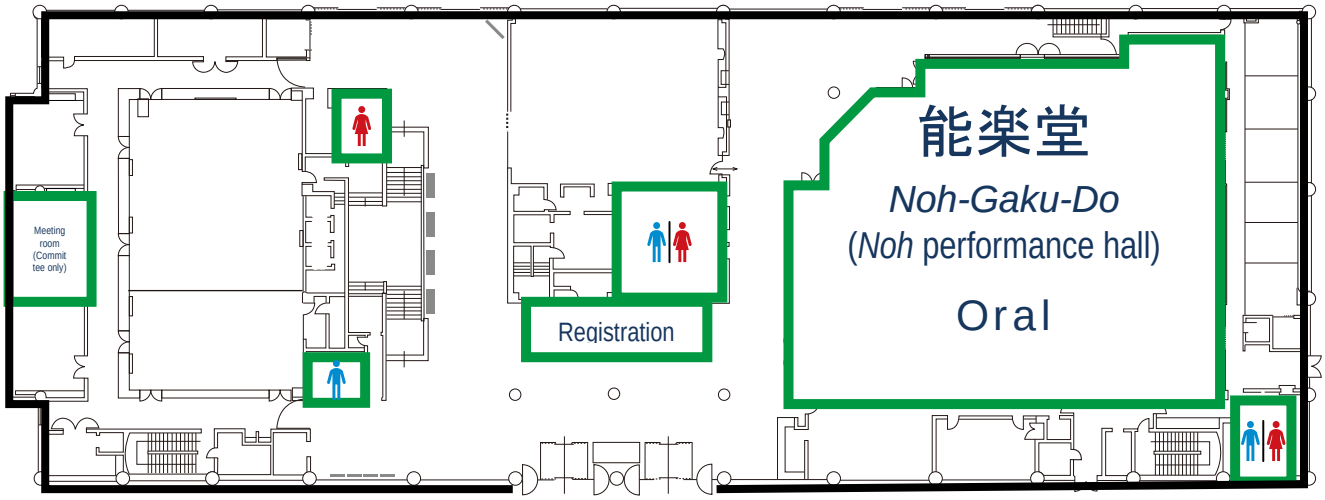
Venue

Nara Kasugano International Forum 甍 IRAKA (101 Kasugano-cho, Nara, 630-8212, Japan)

2F



1F



Inside *Noh-Gaku-Do*, please,

- Do not eat or drink
- Take off shoes on the stage

WiFi

- SSID: **`iraka-free`**
- Password: (none)

Proceedings

Published as Lecture Notes in Computer Science (LNCS), vols. 15520–15524.

Part I Part II Part III Part IV Part V

- Complementary download links will be provided separately during the conference.

Program

Day 1: Wednesday, January 8

9:00 – 17:00 Registration [1F: Entrance Hall]

9:30 – 9:45 **Opening** [1F: *Noh-Gaku-Do*]

9:45 – 10:45 **Keynote Talk 1: Dr. Nancy F. Chen (A*STAR)** [1F: *Noh-Gaku-Do*]
Multimodal, Multilingual Generative AI: From Multicultural Contextualization to Empathetic Reasoning
Chair: Chong-Wah Ngo (Singapore Management University)

«Coffee Break»

11:00 – 12:00 **Best Paper Session** [1F: *Noh-Gaku-Do*]
Chair: Toshihiko Yamasaki (The University of Tokyo)
[196] RoLD: Robot Latent Diffusion for Multi-task Policy Modeling (Tan, Wenhui; Liu, Bei; Zhang, Junbo; Song, Ruihua; Fu, Jianlong)
[379] TDM: Temporally-Consistent Diffusion Model for All-in-One Real-World Video Restoration (Li, Yizhou; Liu, Zihua; Monno, Yusuke; Okutomi, Masatoshi)
[451] ESC-MISR: Enhancing Spatial Correlations for Multi-Image Super-Resolution in Remote Sensing (Zhang, Zhihui; Pang, Jinhui; Li, Jianan; Hao, Xiaoshuai)
[462] Flat Local Minima for Continual Learning on Semantic Segmentation (Huang, Zhongzhan; Liang, Mingfu; Liang, Senwei; Zhong, Shanshan)

«Lunch Break»

14:00 – 15:30 **Poster Session 1** [2F: Reception Hall]
(See pp.7–8 for details)
«Coffee Break»

15:30 – 16:30 **Oral Session 1: Content Generation** [1F: *Noh-Gaku-Do*]
Chair: Luwei Zhang (The University of Tokyo)
[268] AD2AT: Audio Description to Alternative Text, a Dataset of Alternative Text from Movies (Lincker, Elise; Guinaudeau, Camille; Satoh, Shin'ichi)
[310] KuzushijiDiffuser: Japanese Kuzushiji Font Generation with FontDiffuser (Yuan, Honghui; Yanai, Keiji)
[167] Saliency Guided Optimization of Diffusion Latents (Wang, Xiwen; Zhou, Jizhe; Li, Mao; Zhu, Xuekang; Li, Cheng)
[308] Skin-Adapter: Fine-Grained Skin-Color Preservation for Text-to-Image Generation (Chen, Zhuowei; Huang, Mengqi; Chen, Nan; Mao, Zhendong)

«Coffee Break»

16:45 – 17:45 **Oral Session 2: Audio Analysis** [1F: *Noh-Gaku-Do*]

Chair: Ling Xiao (The University of Tokyo)

[273] Operatic Singing Voice Synthesis from Inexperienced Voice Considering Tempo and Vowel Change (Sugahara, Aoto; Kishimoto, Soma; Adachi, Yuji; Tai, Kiyoto; Takashima, Ryoichi; Takiguchi, Tetsuya)

[129] Small Tunes Transformer: Exploring Macro & Micro-Level Hierarchies for Skeleton-Conditioned Melody Generation (Lv, Yishan; Luo, Jing; Ju, Boyuan; Yang, Xinyu)

[430] WavFusion: Towards wav2vec 2.0 Multimodal Speech Emotion Recognition (Li, Feng; Luo, Jiusong; Xia, Wanjun)

[374] SPLGAN-TTS: Learning Semantic and Prosody to Enhance the Text-to-Speech Quality of Lightweight GAN Models (Chang, Ding-Chi; Li, Shiou-Chi; Huang, Jen-Wei)

14:30 – 17:30 Video Browser Showdown (Expert Session) [2F: Reception Hall]

18:00 – 20:00 **Welcome Reception /** [2F: Reception Hall]
Video Browser Showdown (Team Presentation / Novice Session)
(See pp.12–13 for details)

Day 2: Thursday, January 9

9:00 – 16:30 Registration [1F: Entrance Hall]

9:30 – 10:30 **Oral Session 3: Object Detection, Recognition, and Tracking**

Chair: Wei-Ta Chu (National Cheng Kung University) [1F: *Noh-Gaku-Do*]

[236] MineTinyNet-YOLO: An Efficient Small Object Detection Method for Complex Underground Coal Mine Scenarios (Yaling, Hao; Wei, Wu)

[436] Mix-YOLONet: Deep Image Dehazing for Improving Object Detection (Lim, Xin; Wong, Lai-Kuan; Loh, Yuen Peng; Gu, Ke; Lin, Weisi)

[411] Counting Unique Objects in Geo-Tagged Street Images: A Case Study of Homeless Encampments in Los Angeles (Ghasemi, Narges; Kim, Seon Ho; Alfarrarjeh, Abdullah; Shahabi, Cyrus)

[181] HCV: Lightweight Hybrid CNN-Vision Transformer for Visual Object Tracking (Chen, Liang-Chia; Chu, Wei-Ta)

«Coffee Break»

10:45 – 11:30 **Oral Session 4: Trusted and Explainable AI** [1F: *Noh-Gaku-Do*]

Chair: Kazuaki Nakamura (Tokyo University of Science)

[74] Detoxification of Unlabeled Dataset: Reducing Implicit Class Imbalance Using Pseudo-Jacobian of GAN’s Generator (Suyama, Kosei; Nakamura, Kazuaki)

[244] Making Strides Security in Multimodal Fake News Detection Models: A Comprehensive Analysis of Adversarial Attacks (Si, Jiahua; Wang, Youze; Hu, Wenbo; Liu, Qiang; Hong,

Richang)

[415] AMPLE: Emotion-Aware Multimodal Fusion Prompt Learning for Fake News Detection (Xu, Xiaoman; Li, Xiangrun; Wang, Taihang; Jiang, Ye)

«Lunch Break»

13:30 – 15:00 **Poster Session 2 / Demonstrations** [2F: Reception Hall]

(See pp.8–10 / 13–15 for details)

«Coffee Break»

15:00 – 15:45 **Oral Session 5: Signal Processing** [1F: *Noh-Gaku-Do*]

Chair: Masahiro Toyoura (University of Yamanashi)

[297] Uncertainty-Guided Joint Semi-Supervised Segmentation and Registration of Cardiac Images (Chen, Junjian; Yang, Xuan)

[337] Wavelet Integrated Convolutional Neural Network for ECG Signal Denoising (Terada, Takamasa; Toyoura, Masahiro)

[392] MPPQNet: A Moment-Preserving Product Quantization Neural Network for Progressive 3D Point Cloud Transmission (Cheng, Shyi-Chyi; Chen, Yen-Lin; Li, Shih-Yu)

«Coffee Break»

16:00 – 17:00 **Keynote Talk 2: Prof. Kiyoharu Aizawa (The University of Tokyo)** [1F: *Noh-Gaku-Do*]

Manga109 and MangaUB: How Far Can Large Multimodal Models (LMMs) Go in Understanding Manga?

Chair: Keiji Yanai (The University of Electro-Communications)

18:00 – 20:30 **Banquet** [KOTOWA Nara Koen Premium View]

Please expect at least 20 minutes' walk from the conference venue.

Day 3: Friday, January 10

9:00 – 16:30 Registration [1F: Entrance Hall]

9:15 – 10:15 **Keynote Talk 3: Dr. Andrei Bursuc (valeo.ai / INRIA)** [1F: *Noh-Gaku-Do*]

Multi-Modal Foundation Models in the Automotive Industry

Chair: Ichiro Ide (Nagoya University)

«Coffee Break»

10:30 – 11:30 **Oral Session 6: Recognition and Reasoning** [1F: *Noh-Gaku-Do*]

Chair: Satoshi Yamasaki (NEC)

[218] A Multi-Expert Collaborative Framework for Multimodal Named Entity Recognition (Xu, Bo; Jiang, Haiqi; Wei, Shouang; Du, Ming; Song, Hui; Wang, Hongya)

[266] SSDL: Sensor-to-Skeleton Diffusion Model with Lipschitz Regularization for Human Activity Recognition (Sharma, Nikhil; Sun, Changchang; Zhao, Zhenghao; Ngu, Anne Hee Hiong; Latapie, Hugo; Yan, Yan)

[395] Open-Vocabulary Scene Graph Generation via Synonym-Based Predicate Descriptor (Goto, Yuta; Yamazaki, Satoshi; Shibata, Takashi; Liu, Jianquan)

[274] Grounding Deliberate Reasoning in Multimodal Large Language Models (Chen, Jiaxing; Liu, Yuxuan; Li, Dehu; An, Xiang; Deng, Weimo; Feng, Ziyong; Zhao, Yongle; Xie, Yin)

«Lunch Break»

13:30 – 15:00 **Poster Session 3 / Demonstrations** [2F: Reception Hall]

(See pp.10–12 / 13–15 for details)

«Coffee Break»

15:00 – 16:00 **Special Session MLLMA Oral Session** [1F: *Noh-Gaku-Do*]

—Multimodal Large Language Models and Applications—

Chair: Rajiv Ratn Shah (IIIT-Delhi)

[193] Image2Text2Image: A Novel Framework for Label-Free Evaluation of Image-to-Text Generation with Text-to-Image Diffusion Models (Huang, Jia-Hong; Zhu, Hongyi; Shen, Yixian; Rudinac, Stevan; Kanoulas, Evangelos)

[288] Enhanced Anomaly Detection in 3D Motion through Language-Inspired Occlusion-Aware Modeling (Li, Su; Wang, Liang; Wang, Jianye; Zhang, Ziheng; Zhang, Junjun; Zhang, Lei)

[364] Evaluating VQA Models' Consistency in the Scientific Domain (C. Quan, Khanh-An; Guinaudeau, Camille; Satoh, Shin'ichi)

Panel Discussion

«Coffee Break»

16:15 – 17:00 **Oral Session 7: Search and Retrieval** [1F: *Noh-Gaku-Do*]

Chair: Nicolas Michel (The University of Tokyo)

[346] RobSparse: Automatic Search for GPU-Friendly Robust and Sparse Vision Transformers (Su, Yulan; Zhang, Sisi; Wang, Yan; Wang, Xingbin; Zhao, Lutan; Dan, Meng; Hou, Rui)

[232] Image-Generation AI Model Retrieval by Contrastive Learning-Based Style Distance Calculation (Vu, Thi Ngoc Anh; Shoji, Yoshiyuki; Oe, Yuma; Pham, Huu Long; Ohshima,

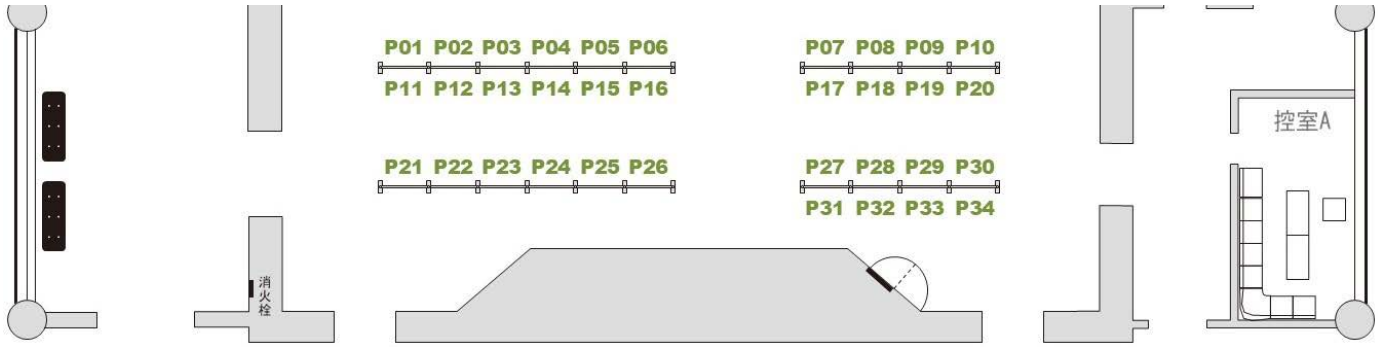
Hiroaki)

[414] Dynamic Exploration Graph: A Novel Approach for Efficient Nearest Neighbor Search in Evolving Multimedia Datasets (Hezel, Nico; Barthel, Kai Uwe; Schilling, Bruno; Schall, Konstantin; Jung, Klaus)

17:00 – 17:15 **Closing** [1F: *Noh-Gaku-Do*]

Posters

[2F: Reception Hall]



- Digital posters (Those labeled as DP) will be presented on the display.
- Posters for oral papers presented on each day can be presented at P24–P34.

Day 1: Poster Session 1 (PS1)

[P01 (120)] Quantized-ViT Efficient Training via Fisher Matrix Regularization (Shang, Yuzhang; Liu, Gaowen; Kompella, Ramana; Yan, Yan)

[P02 (121)] Saliency Based Data Augmentation for Few-Shot Video Action Recognition (Kong, Yongqiang; Wang, Yunhong; Li, Annan)

[P03 (128)] Hybrid Scalable Video Coding with Neural Compression and Enhancement for Streaming Media (Ye, Yuyao; Yang, Jiayu; Zhao, Yang; Gao, Mengping; Cao, Hongbin; Wang, Ronggang)

[P04 (130)] Pubic Symphysis-Fetal Head Segmentation Network Using BiFormer Attention Mechanism and Multipath Dilated Convolution (Cai, Pengzhou; Jiang, Lu; Li, Yanxin; Liu, Xiaojuan; Lan, Libin)

[P05 (131)] DART: Depth-Enhanced Accurate and Real-Time Background Matting (Li, Guofeng; Li, Hanxi; Li, Bo; Wu, Lin; Cheng, Yan)

[P06 (141)] MLP-AMDC: A MLP Architecture for Adaptive-Mask-based Dual-Camera Snapshot Hyperspectral Imaging (Cai, Zeyu; Chen, Xunhao; Zhang, Can; Chen, yuchong; Yang, Jiming; Shi, Wubin; Jin, Chengqian; Da, Feipeng)

[P07 (144)] Kiite World: Socializing Map-Based Music Exploration Through Playlist Sharing and Synchronized Listening (Tsukuda, Kosetsu; Takahashi, Takumi; Ishida, Keisuke; Hamasaki, Masahiro; Goto, Masataka)

[P08 (146)] Enhancing Environmental Monitoring through Multispectral Imaging: The WasteMS Dataset for Semantic Segmentation of Lakeside Waste (Zhu, Qinfeng; Weng, Ningxin; Fan, Lei; Cai, Yuanzhi)

[P09 (158)] Frequency-Aware Convolution for Sound Event Detection (Song, Tao; Zhang, Wenwen)

[P10 (163)] MSD-YOLO: An Efficient Algorithm for Small Target Detection (Liu, Dongyu; Zhu, Yuan; liu, rui; Xing, Zhecong; Geng, Weiyang; Wang, Yanqiang)

[P11 (166)] Robust Active Speaker Detection in Challenging Environments Using GNN-Fused Multi-Modal Cues and Body Language (Li, Yongqian; Luo, Yong; Zhou, Xin)

[P12 (172)] Intra-Class Compact Facial Expression Recognition Based on Amplitude Phase Separation (Tian, Xiang; Zhang, Yuan; Mu, Chang; Zhang, Ziyang)

[P13 (176)] PA2Net: Pyramid Attention Aggregation Network for Saliency Detection (Yu, Jizhe; Liu, Yu; Wu, Xiaoshuai; Xu, Kaiping; Li, Jiangquan)

[P14 (188)] LIESA: Low-Light Image Enhancement with Semantic Awareness (Zhang, Jingyao; Hao, Shijie; Sun, Fuming Sun; Rao, Yuan)

[P15 (195)] Deep Dual Internal Learning for Hyperspectral Image Super-Resolution (Sun, Yongqing; Liu, Hong; Chang, Qiong; Han, Xianhua)

[P16 (198)] Zero-Shot Sketch-Based Image Retrieval with Hybrid Information Fusion and Sample Relationship Modeling (Wu, Weijie; Li, Jun; Wu, Zhijian; Xu, Jianhua)

[P17 (206)] The Right to an Explanation under the GDPR and the AI Act (Juliussen, Bjørn Aslak)

[P18 (221)] Improving Singing Voice Transcription Generalization with AI Generated Accompaniments (Perez, Miguel; Kirchhoff, Holger; Grosche, Peter; Serra, Xavier)

[P19 (228)] LITA: LMM-Guided Image-Text Alignment for Art Assessment (Sunada, Tatsumi; Shiohara, Kaede; Xiao, Ling; Yamasaki, Toshihiko)

[P20 (229)] Towards Inclusive Education: Multimodal Classification of Textbook Images for Accessibility (Yadav, Saumya; Lincker, Élise; Huron, Caroline; Martin, Stéphanie; Guinaudeau, Camille; Satoh, Shin’ichi; Shukla, Jainendra)

[P21 (296)] GWUNet: A UNet with Gated Attention and Improved Wavelet Transform for Thyroid Nodules Segmentation (Zheng, Shuijing; Yu, Suxi; Wang, Yi; Wen, Jing)

[P22 (111)] SCLSTE: Semi-Supervised Contrastive Learning-Guided Scene Text Editing (Yin, Min; Xie, Liang; Liang, HaoRan; Zhao, Xing; Chen, Ben; Liang, RongHua)

[P23 (237)] Hyper-NeuS: Hypernetworks for Neural SDF Implicit Surface Reconstruction by Volume Rendering (Li, Jingkun; Qi, Na; Zhu, Qing)

Day 2: Poster Session 2 (PS2)

[P01 (192)] Comparative Analysis of Relevance Feedback Techniques for Image Retrieval (Vadicamo, Lucia; Scotti, Francesca; Dearle, Alan; Connor, Richard)

[P02 (241)] Understanding the Roles of Visual Modality in Multimodal Dialogue: An Empirical Study (Cao, Qian; Song, Ruihua; Chen, Xu)

[P03 (242)] DistillSleep: Leverage Self-Distillation to Improve Performance After Representation Learning for Sleep Staging (Yu, Le; Zhang, Xianchao; Qian, Shuxia; Sun, Hong)

[P04 (246)] Temporal Closeness for Enhanced Cross-Modal Retrieval of Sensor and Image Data (Yamamoto, Shuhei; Kando, Noriko)

[P05 (247)] An Analytical Method for Rendering Plenoptic Cameras 2.0 on 3D Multi-Layer Displays (Losfeld, Armand; Seznec, Nicolas; Van Bogaert, Laurie; Lafruit, Gauthier; Teratani, Mehrdad)

[P06 (251)] QRALadder: QoE and Resource Consumption-Aware Encoding Ladder Optimization for Live Video Streaming (Zhu, Yingqian; Gao, Guanyu)

[P07 (256)] Boosting Human Pose Estimation via Heatmap Refinement (Jiang, Ling; Liu, Zhuocheng; Li,

Kaige; Wu, Wei)

[P08 (265)] FoodMLLM-JP: Leveraging Multimodal Large Language Models for Japanese Recipe Generation (Imajuku, Yuki; Yamakata, Yoko; Aizawa, Kiyoharu)

[P09 (283)] LLMs-Based Augmentation for Domain Adaptation in Long-Tailed Food Datasets (Wang, Qing; Ngo, Chong Wah; Lim, Ee-Peng; Sun, Qianru)

[P10 (292)] Music2MIDI: Pop Music to MIDI Piano Cover Generation (Yip, Tin Yui; Chau, Chuck-Jee)

[P11 (293)] Balancing Efficiency and Accuracy: An Analysis of Sampling for Video Copy Detection (Chen, Xiangyu; Satoh, Shinichi)

[P12 (295)] One-Shot Generative Domain Adaptation by Constructing Self-Amplifying Datasets (Xiang, Yanru; Li, Yi)

[P13 (306)] Visual Anomaly Detection on Topological Connectivity under Improved YOLOv8 (Li, Yu; Xie, Zhenping)

[P14 (315)] HierArtEx: Hierarchical Representations and Art Experts Supporting the Retrieval of Museums in the Metaverse (Falcon, Alex; Abdari, Ali; Serra, Giuseppe)

[P15 (317)] DocMamba: Robust Document Image Dewarping via Selective State Space Sequence Modeling (Han, Miaolin; Li, Huibin)

[P16 (326)] Real-Time Action Detection in Volleyball Matches Using DETR Architecture (Shih, Mu-Jan; Hsu, Yi-Yu)

[P17 (332)] Select and Order: Enhancing Few-Shot Image Classification Through In-Context Learning (Huang, Hujiang; Xie, Yu; Gao, Jun; Fan, Chuanliu; Cao, Ziqiang)

[P18 (336)] SMG-Diff: Adversarial Attack Method Based on Semantic Mask-Guided Diffusion (Zhang, Yongliang; Liu, Jing)

[P19 (344)] Dual-Task Feedback Learning for Tongue Detection via Super-Resolution Integration (Sun, Ying; Wei, Meiyi; Chen, Gang)

[P20 (354)] Towards Visual Storytelling by Understanding Narrative Context through Scene-Graphs (Phueaksri, Itthisak; Kastner, Marc A.; Kawanishi, Yasutomo; Komamizu, Takahiro; Ide, Ichiro)

[P21 (456)] AMFT-YOLO: A Adaptive Multi-Scale YOLO Algorithm with Multi-Level Feature Fusion for Object Detection in UAV Scenes (Wang, Tiebiao; Li, Xiaoyang; Cui, Zhenchao)

[P22 (276)] Lightweight Dual Grouped Large-Kernel Convolutions for Salient Object Detection Network (Liu, Jiajie; Zhang, Zhibin)

[P23 (312)] Modeling High-Order Relationships between Human and Video for Emotion Recognition (Ai, Hanxu; Tao, Xiaomei; Li, Xingbing; Gan, Yanling)

[DP (117)] EIA: Edge-Aware Imperceptible Adversarial Attacks on 3D Point Clouds (Wang, Zhensu; Peng, Weilong; Wang, Le; Wu, Zhizhe; Zhu, Peican; Tang, Keke)

[DP (127)] MKSNet: Advanced Small Object Detection in Remote Sensing Imagery with Multi-Kernel and Dual Attention Mechanisms (Zhang, Jiahao; Gao, Guangyu; Zhao, Xiao)

[DP (140)] Infrared Small Target Detection with Feature Refinement and Context Enhancement (Li, Xiuhong; Zhu, Xinyue; Li, Boyuan; Li, Songlin; Wang, Luyao; Jia, Zhenhong)

[DP (173)] Modality-Specific Hashing: Transform Cross-Modal Retrieval into Single-Modal Retrieval (Ding, Guohui; Li, Zhonghua; Ren, Yongqiang)

[DP (178)] Multimodal Prompt Learning for Audio Visual Scene-aware Dialog (Xu, Feifei; Jia, Fumiaoyue; Zhou, Wang)

[DP (182)] MSA-Former: Multi-Scale Adaptive Transformer for Image Snow Removal (Wang, Bin; Chen, Zekun; Zhang, Lei; Liang, Shili; Guo, Sijia; Kang, Xinyu; Li, Huajing)

[DP (184)] SES-Net: Multi-dimensional Spot-Edge-Surface Network for Nuclei Segmentation (Lu, Congjian; Zhou, Shuwang; Shan, Ke; Zhang, Hongkuan; Liu, Zhaoyang)

[DP (189)] PianoPal: A Robotic Multimedia System for Interactive Piano Instruction Based on Q-Learning and Real-Time Feedback (Wang, Yufei; Yao, Junfeng; Wang, Zefeng)

[DP (199)] CLIP Multi-Modal Hashing for Multimedia Retrieval (Zhu, Jian; Sheng, Mingkai; Huang, Zhangmin; Chang, Jingfei; Long, Jian; Jiang, Jinling; Liu, Lei; Luo, Cheng)

[DP (223)] Integrating S1&S2 Framework for Enhanced Semantic Match in Person Re-Identification (Yang, Xiukang; Ge, Jingguo; Li, Hui; Li, Liangxiong; Wu, Bingzhen)

[DP (253)] Structural Information-Guided Fine-Grained Texture Image Inpainting (Fang, Zhiyi; Qian, Yi; Dai, Xiyue)

[DP (272)] GFA-UDIS: Global-to-Flow Alignment for Unsupervised Deep Image Stitching (Han, Sijia; Zhang, Zhibin)

[DP (275)] Joint Decision Network with Modality-Specific and Dual Interactive Features for Fake News Detection (Wu, Fei; Zhou, Ruixuan; Ji, Yimu; Jing, Xiao-Yuan)

[DP (277)] MS-SAM: Multi-Scale SAM Based on Dynamic Weighted Agent Attention (Yang, Enhui; Zhang, Zhibin)

[DP (281)] Multi-Modal Information Multi-Angle Mining for Multimedia Recommendation (Zhu, Yijie; Li, MingYong)

[DP (305)] MambaTalk: Speech-Driven 3D Facial Animation with Mamba (Zhu, Deli; Xu, Zhao; Yang, Yunong)

Day 3: Poster Session 3 (PS3)

[P01 (356)] Rotation Methods for 360-degree Videos in Virtual Reality —A Comparative Study (Hürst, Wolfgang; Zeches, Leo)

[P02 (360)] Camouflaged Object Detection Based on Localization Guidance and Multi-Scale Refinement (Wang, JinYang; Wu, Wei)

[P03 (362)] Poseidon: A NAS-Based Ensemble Defense Method against Multiple Perturbations (Su, Yulan; Zhang, Sisi; Lin, Zechao; Wang, Xingbin; Zhao, Lutan; Meng, Dan; Hou, Rui)

[P04 (363)] MM-CARP: Multimodal Model with Cross-Modal Retrieval-Augmented and Visual Region Perception (Guo, Junhao; Fu, Chenhan; Wang, Guoming; Lu, Rongxing; Chen, Dong; Tang, Siliang)

[P05 (365)] Revisit Data Association in Semantic SLAM Systems for Autonomous Parking (Shao, Xuan; Huang, Leming; Liu, Xinghua)

[P06 (368)] Lightweight Motion-Aware Video Super-Resolution for Compressed Videos (Kwon, Ilhwan; Li, Jun; Shah, Rajiv Ratn; Prasad, Mukesh)

[P07 (373)] Vision-Language Pretraining for Variable-Shot Image Classification (Papadopoulos, Sotirios; Ioannidis, Konstantinos; Vrochidis, Stefanos; Kompatsiaris, Ioannis; Patras, Ioannis)

[P08 (377)] A Multi-Aspect Multi-Granularity Pronunciation Assessment Method Based on Branchformer Encoder and Hierarchical Aggregation (Du, Wenxu; Wumaier, Aishan; Shi, Yahui; Yi, Nian; Liu, Dehua)

[P09 (386)] SCANet: Semantic Coherence Attention Network for Clothing Change Person Re-

Identification (Yang, Dajiang; Wu, Wei; Lee, Yuxing)

[P10 (417)] Toward a Full Pipeline Approach to Autonomous Drone Landing Site Identification: From Terrain Survey to Embedded Classifier (Springer, Joshua David; Guðmundsson, Gylfi Þór; Kyas, Marcel)

[P11 (429)] Innovative Lifelog Visualization and Exploration in Virtual Reality —A Comparative Study (Hürst, Wolfgang; Visser, Yannick)

[P12 (435)] Synchronization and Calibration of Video Sequences Acquired Using Multiple Plenoptic 2.0 Cameras (Bonatto, Daniele; Fernandes Pinto Fachada, Sarah; Sancho, Jaime; Juarez, Eduardo; Lafruit, Gauthier; Teratani, Mehrdad)

[P13 (444)] A Dual-Branch Model for Color Constancy (Chen, Zhaoxin; Ma, Bo)

[P14 (445)] Data-Free Functional Projection of Large Language Models onto Social Media Tagging Domain (Mu, Wenchuan; Lim, Kwan Hui)

[P15 (455)] MDT-Net: A Mask Decoder Tuning Strategy for CLIP-Based Zero-Shot 3D Classification (Yan, Hao; Bai, Jing)

[P16 (458)] Optimally Planning Drone Trajectory to Capture a 3D Gaussian Splatting Object (Wu, Cheng-Yuan; Sun, Yuan-Chun; Lee, Cheng-Tse; Hsu, Cheng-Hsin)

[P17 (230)] Quantifying Image-Adjective Associations by Leveraging Large-Scale Pretrained Models (Matsuhira, Chihaya; Kastner, Marc A.; Komamizu, Takahiro; Hirayama, Takatsugu; Ide, Ichiro)

[P18 (137)] Can Masking Background and Object Reduce Static Bias for Zero-Shot Action Recognition? (Fukuzawa, Takumi; Hara, Kensho; Kataoka, Hirokatsu; Tamaki, Toru)

[P19 (355)] CalorieVoL: Integrating Volumetric Context into Multimodal Large Language Models for Image-Based Calorie Estimation (Tanabe, Hikaru; Yanai, Keiji)

[P20 (416)] Multimodal Engagement Prediction in Human-Robot Interaction using Transformer Neural Networks (Lim, Jia Yap; See, John; Dondrup, Christian)

[P21 (431)] What Should Autonomous Robots Verbalize and What Should They Not? (Yoshihara, Daichi; Yuguchi, Akishige; Kawano, Seiya; Iio, Takamasa; Yoshino, Koichiro)

[P22 (438)] BiCA-YOLO: Bidirectional Feature Enhancement and Cross Coordinate Attention for Small Object Detection (Lv, Jinyan; Xiao, Guoqiang)

[DP (307)] Frequency-Based Unsupervised Low-Light Image Enhancement Framework (Wang, Haodian)

[DP (309)] Target-Oriented Dynamic Denosing Curriculum Learning for Multimodel Stance Detection (Suo, Zihao; Pan, Shanliang)

[DP (316)] Noise-Robust Separating Multi-Source Aliased Vibration Signal Based on Transformer Demucs (Jiang, Wanchang; Jiang, Yuxin)

[DP (321)] gFlow: Distributed Real-Time Reverse Remote Rendering System Model (Xu, Yixiao; Li, Yubo; Xu, Wanzhao; Gu, Yicheng; Wang, Yun; Ma, Jiangyuan; Qi, Zhengwei)

[DP (331)] BLCC: A Benchmark for Multi-LiDAR and Multi-Camera Calibration (Minghui, Hou; Gang, Wang; Zhiyang, Wang; Tongzhou, Zhang; Baorui, Ma)

[DP (342)] MC-YOLO: Multi-Scale Transmission Line Defect Target Recognition Network (Wang, Jingdong; Ding, Xu; Meng, Fanqi)

[DP (350)] A Novel Human Abnormal Posture Detection Method Based on Spatial-Topological Feature Fusion of Skeleton (Ma, Yuefeng; Cheng, Zhiqi; Liu, Deheng; Tang, Shiying)

[DP (359)] SSCDUF: Spatial-Spectral Correlation Transformer Based on Deep Unfolding Framework for

Hyperspectral Image Reconstruction (Zhao, Hui; Qi, Na; Zhu, Qing; Lin, Xiumin)

[DP (383)] Cross-View Geo-Localization via Learning Correspondence Semantic Similarity Knowledge (Chen, Guanli; Huang, Guoheng; Yuan, Xiaochen; Chen, Xuhang; Zhong, Guo; Pun, Chi-Man)

[DP (385)] Style Separation and Content Recovery for Generalizable Sketch Re-Identification and a New Benchmark (Lu, Lingyi; Xu, Xin; Wang, Xiao)

[DP (387)] Chain of Thought Guided Few-Shot Fine-Tuning of LLMs for Multimodal Aspect-Based Sentiment Classification (Wu, Hao; Yang, Danping; Liu, Peng; Li, Xianxian)

[DP (393)] Progressive Neural Architecture Generation with Weaker Predictors (Zhang, Zhengzhuo; Zhuang, Liansheng)

[DP (420)] Self-Supervised Reference-based Image Super-Resolution with Conditional Diffusion Model (Shi, Shuai; Qi, Na; Li, Yezi; Zhu, Qing)

[DP (447)] TPS-YOLO: The Efficient Tiny Person Detection Network Based on Improved YOLOv8 and Model Pruning (Yao, Li; Huang, Qianni; Wan, Yan)

[DP (460)] MICAN: Multi-Modal Inconsistency-Based Cooperation Attention Network for Fake News Detection (Yi, Zepu; Lu, Songfeng; Tang, Xueming; Zhu, Jianxin; Wu, Junjun)

[DP (214)] TACST: Time-Aware Transformer for Robust Speech Emotion Recognition (Wei, Wei; Zhang, Bingkun; Wang, Yibing)

[DP (215)] TS-MEFM: A New Multimodal Speech Emotion Recognition Network Based on Speech and Text Fusion (Wei, Wei; Zhang, Bingkun; Wang, Yibing)

Video Browser Showdown (Day 1) [2F: Reception Hall]

[406] VEAGLE: Eye Gaze-Assisted Guidance for Video Browser Showdown (Nguyen-Ho, Thang-Long; Huynh, Viet-Tham; Kongmeesub, Onanong; Tran, Minh-Triet; Nie, Dongyun; Healy, Graham; Gurrin, Cathal)

[501] VideoEase at VBS2025: An Interactive Video Retrieval System (Tran, Quang-Linh; Nguyen, Binh; Jones, Gareth J. F.; Gurrin, Cathal)

[502] Feature-Driven Video Segmentation and Advanced Querying with vitrivr-engine (Rossetto, Luca; Gasser, Ralph)

[503] HORUS: Multimodal Large Language Models Framework for Video Retrieval at VBS 2025 (Nguyen, Tai; Vo, Anh Ngoc Minh; Pham, Dat Duc; Tran, Vinh Quang; Duong, Nhu Thi Quynh; Le, Tien Anh; Le, Tan Duy; Nguyen, Binh T.)

[504] Interactive Video Search with Multi-Modal LLM Video Captioning (Cheng, Yu Tong; Wu, Jiaxin; Ma, Zhixin; He, Jiangshan; Wei, Xiao-Yong; Ngo, Chong Wah)

[505] FUSIONISTA: Fusion of 3-D Information of Video in Retrieval System (Le, Huy M.; Nguyen Tien, Dat; Le Duy, Khang; Nguyen Dang Quang, Tuan; Nguyen Khanh, Toan; Nguyen, Binh T.)

[506] ViFi: A Video Finding System at Video Browser Showdown 2025 (C. Quan, Khanh-An; Ngoc Nguyen, Qui; Tran, Minh-Triet)

[507] ViewsInsight2.0: Enhancing Video Retrieval for VBS 2025 with an Automatic Query Generator Powered by Large Language Models (Vuong, Gia-Huy; Ho, Van-Son; Nguyen-Dang, Tien-Thanh; Thai, Xuan-Dang; Ho-Le, Minh-Quan; Le, Tu-Khiem; Pham, Minh-Khoi; Ninh, Van-Tu; Gurrin, Cathal; Tran, Minh-Triet)

[508] VERGE in VBS 2025 (Pantelidis, Nick; Georgalis, Dimitris; Pegia, Maria; Galanopoulos, Damianos;

Apostolidis, Konstantinos; Stavrothanasopoulos, Klearchos; Moumtzidou, Anastasia; Gkountakos, Konstantinos; Gialampoukidis, Ilias; Vrochidis, Stefanos; Mezaris, Vasileios; Kompatsiaris, Ioannis)

[509] Exquisitor at the Video Browser Showdown 2025: Unifying Conversational Search and User Relevance Feedback (Sharma, Ujjwal; Khan, Omar Shahbaz; Rudinac, Stevan; Jónsson, Björn Þór)

[510] Simplified Video Retrieval in Virtual Reality with vitrivr-VR (Spiess, Florian; Rossetto, Luca; Schuldt, Heiko)

[511] diveXplore at the Video Browser Showdown 2025 (Leopold, Mario; Schöffmann, Klaus)

[512] NII-UIT at VBS2025: Multimodal Video Retrieval with LLM Integration and Dynamic Temporal Search (Tran Gia, Bao; Bui Cong Khanh, Tuong; Le Thi Thanh, Tam; Tran Doan, Thuyen; Le Tran Trong, Khiem; Do, Tien; Mai, Tien-Dung; Duc Ngo, Thanh; Le, Duy-Dinh; Satoh, Shin'ichi)

[513] PraK Tool V3: Enhancing Video Item Search Using Localized Text and Texture Queries (Stroh, Michael; Kloda, Vojtěch; Verner, Benjamin; Vopálková, Zuzana; Buchmüller, Raphael; Jäckl, Bastian; Lokoč, Jakub; Hajko, Jakob)

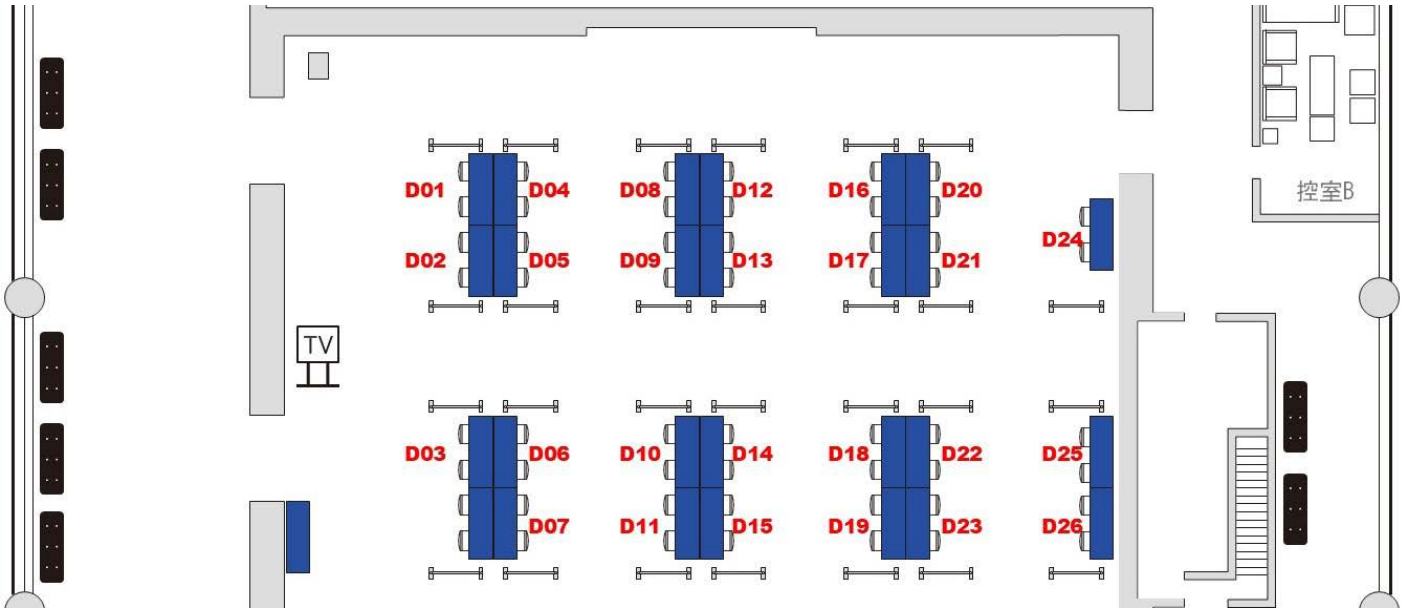
[514] MediaMix: Multimedia Retrieval in Mixed Reality (Arnold, Rahel; Kempf, Rahel; Waltenspül, Raphael; Schuldt, Heiko)

[515] SnapSeek 2.0 at Video Browser Showdown 2025 (Ho-Le, Minh-Quan; Ho, Duy-Khang; Do-Huu, Huy-Hoang; Le-Hinh, Nhut-Thanh; Vo-Hoang, Hoa-Vien; Ninh, Van-Tu; Gurrin, Cathal; Tran, Minh-Triet)

[517] IMSearch 2.0: Toward User-Centric and Efficient Interactive Multimedia Retrieval System (Luu, Duc-Tuan; C. Quan, Khanh-An; Nguyen, Duy-Ngoc; Bui-Le, Khanh-Linh; Doan, Nhat-Sang; Le-Ngo, Minh-Duc; Nguyen, Vinh-Tiep; Tran, Minh-Triet)

Demonstrations (Days 2 & 3)

[2F: Reception Hall]



[D01 (468)] SelectSum: Topic-Based Selective Summarization of Speech-Based Videos (Wattasseril, Jobin Idiculla; Döllner, Jürgen)

[D02 (469)] Real-Time Visualizer for Turntablist Performance (Hamanaka, Masatoshi)

[D03 (494)] Multi-Dimensional Exploration of Media Collection Metadata (Khan, Omar Shahbaz; Duane, Aaron; Hasnan, Hariz; Blavec, Noé Le; Ouvrard, Pierre; Verdon, Johan; d'Orazio, Laurent; Thierry, Constance; Jónsson, Björn Þór)

[D04 (470)] DriveCoach: Smart Driving Assistance with Multimodal Risk Prediction and Risk Adaptive Behavior Recommendation (Gan, Wenbin; Dao, Minh-Son; Zettsu, Koji)

[D05 (472)] System Demo of Modeling Smart University Campus Virtual Environments (Fernandez Roblero, Jaime Boanerjes ; Ali, Muhammad Intizar)

[D06 (473)] AMDA: Advancing Multimedia Data Annotation for Human-Centric Situations (Mohamed Serouis, Ibrahim; Sèdes, Florence)

[D07 (475)] FencBuddy: Action-Aware Depth Perception Training for Fencing Attacks (Hung-Yao, Peng; Zi-Heng, Zhong; Cheng-Chih, Tsai; Ching-Yeh, Chiang; Tse-Yu, Pan)

[D08 (477)] WaveFontStyler: Font Style Transfer Based on Sound (Izumi, Kota; Yanai, Keiji)

[D09 (479)] Training a Segmentation-Based Visual Anonymization Service for Street Scenes (Korb, Martin; Bailer, Werner)

[D10 (481)] CleverFox: Integrating Visual Mnemonics with AI for Enhanced Language Learning (Chiang, Yung-Chu; Tang, Zi-Xian; Luo, Yi-Ching; Chang, Jason S.)

[D11 (482)] Fingering Prediction for Classical Guitar: Dataset Creation and Model Development (Iino, Nami; Iino, Akinaru)

[D12 (483)] An Implementation of Networked JamSketch (Kitahara, Tetsuro; Tsutsumi, Takuya; Nagoshi, Takaaki; Suzuki, Taizan)

[D13 (485)] Using Language Models to Generate and Forget the Narrative Memories of an Assistive Robot (Garcia Contreras, Angel Fernando; Chang, Wen-Yu; Kawano, Seiya; Chen, Yun-Nung; Yoshino, Koichiro)

[D14 (486)] Better Image Segmentation with Classification: Guiding Zero-Shot Models Using Class Activation Maps (Borgli, Hanna; Stensland, Håkon Kvale; Halvorsen, Pål)

[D15 (488)] Transformer-Based Audio Generation Conditioned by 2D Latent Maps: A Demonstration (Limberg, Christian; Zhang, Zhe; Kastner, Marc A.)

[D16 (489)] KuzushijiFontDiff: Diffusion Model for Japanese Kuzushiji Font Generation (Yuan, Honghui; Yanai, Keiji)

[D17 (490)] SceneTextStyler: Editing Text with Style Transformation (Yuan, Honghui; Yanai, Keiji)

[D18 (492)] Multimodal Interoperability with the CLAMS Platform (Lynch, Kelley; Rim, Kyeongmin; King, Owen; Pustejovsky, James)

[D19 (493)] Enhancing User Control in AI-Based Video Summarization for Social Media (Kontostathis, Ioannis; Apostolidis, Evlampios; Apostolidis, Konstantinos; Mezaris, Vasileios)

[D20 (496)] Movie Retrieval Systems Using Genre-Guided Multimodal Learning Techniques (Huang, Wei-Lun; Hidayati, Shintami Chusnul; Pan, Tse-Yu)

[D21 (497)] A User Identification and Reading Style Detection System Based on Eye Movement Patterns During Reading (Kongmeesub, Onanong; Gurrin, Cathal; Nie, Dongyun)

[D22 (484)] Federated Learning with Multimodal-Sensing and Knowledge Distillation: An Application on Real-World Benchmark Dataset (Le, Duy-Dong; Huynh, Duy-Thanh; Bao, Pham The)

[D23 (499)] Efficient Deployment of Multimodal AI Models: Leveraging Pruning, Quantization and Multi-Objective Optimization for Edge Computing (Vu, Dang; Dang, Tien; Nguyen, Quoc-Trung; Pham, Tan)

[D24 (466)] Badminton Footwork Practice via an Immersive Virtual Reality System (Jheng, Duen-Chian; Harchan, Bill Louis; Kostka de Sztemberg, Berenika Nawoja; Hsu, Jen-Hao; Hu, Min-Chun)

[D25 (480)] RoboDJ: Live Commentary Robots System Driven by Physical- and Cyber-World

Observations (Kawanishi, Yasutomo; Nakamura, Yutaka; Shintani, Taiken; Ishi, Carlos T.; Kawano, Seiya; Yoshino, Koichiro; Minato, Takashi; Minoh, Michihiko)

[D26 (487)] Leveraging Latent Diffusion in 3D Gaussian Splatting for Novel View Synthesis (Li, Bohan; Li, Xingyi; Liang, Yangwen; Wang, Shuangquan; Song, Kee-Bong)

Voting Instructions for the Demonstration Award

The demonstration award is decided based on the voting by **on-site participants (You!!!)**. Please,

- Find the ballot inside your name tag holder.
- Vote for **up to three** good demonstrations excluding those from your own affiliation (University, Institute, …).
- Post it in the ballot box in the **Reception Hall** by **15:00** on **Day 3**.


Ballot box

Presenting Instructions

Oral Presenters

- Time slot: 12 minutes presentation + 3 minutes Q&A
- Please come to the stage in the *Noh-Gaku-Do* (1F) with your PC, **15 minutes before** the session starts.
- Please take your **shoes off** on the stage.
- Posters can also be presented in the afternoon on the same day. Please use any poster board at P24–P34.

Poster Presenters

- Please set up your poster **after 13:00** and before your session starts on the day of your presentation.
- Please **be available** for discussion in front of your poster **throughout the session**.
- Please remove your poster after,
  - The Reception (for Day 1).
  - The session ends (for Days 2 & 3).

VBS Participants

- Preparation: 13:00–14:30 on Day 1
  - If you need additional time in the morning, please contact the registration desk.
- Please wrap everything up after the Reception before 20:30.

Demonstrators

- Preparation: 9:30–13:30 on Day 2

- Please **be available** for discussion at your demonstration booth **throughout the sessions** on Days 2 & 3.
- Please wrap up everything before 18:00 on Day 3.

External Support

Sponsors

Springer

一般財団法人
奈良県ビジターズビューロー
Nara Visitors Bureau

Subsidies

NICT National Institute of Information and Communications Technology

公益財団法人
電気通信普及財団
The Telecommunications Advancement Foundation

公益財団法人
立石科学技術振興財団

Supporters

IEICE

IPS Japan

ITE