Teaser Session T1

Paper id	Paper title
mmfp2071	Hierarchical Category-Enhanced Prototype Learning for Imbalanced Temporal Recommendation
mmfp2187	Guided Image Synthesis via Initial Image Editing in Diffusion Model
mmfp2273	Tran-GCN: Multi-label Pattern Image Retrieval via Transformer Driven Graph Convolutional Network
mmfp2312	PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes using Large Language Models
mmfp2373	SAUNet: Spatial-Attention Unfolding Network for Image Compressive Sensing
mmfp2527	EmotionKD: A Cross-Modal Knowledge Distillation Framework for Emotion Recognition Based on Physiological Signals
mmfp2588	SAAML: A Framework for Semi-supervised Affective Adaptation via Metric Learning
mmfp2860	In-processing User Constrained Dominant Sets for User-Oriented Fairness in Recommender Systems
mmfp3002	Enhancing Vision-Language Pre-Training with Jointly Learned Questioner and Dense Captioner
mmfp3136	Making Users Indistinguishable: Attribute-wise Unlearning in Recommender Systems
mmfp3332	Pretrained Implicit-Ensemble Transformer for Open-Set Authentication on Multimodal Mobile Biometrics
mmfp3372	Improving Image Captioning through Visual and Semantic Mutual Promotion
mmfp3576	Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection
mmfp3687	Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation
mmfp3691	Efficient Labelling of Affective Video Datasets via Few-Shot & Multi-Task Contrastive Learning
mmfp3733	Equivariant Learning for Out-of-Distribution Cold-start Recommendation
mmfp3743	Globally-Robust Instance Identification and Locally-Accurate Keypoint Alignment for Multi-Person Pose Estimation
mmfp3748	Training Multimedia Event Extraction With Generated Images and Captions
mmfp3771	Understanding User Behavior in Volumetric Video Watching: Dataset, Analysis and Prediction
mmfp4087	Learning from Easy to Hard Pairs: Multi-step Reasoning Network for Human-Object Interaction Detection
mmfp4224	COVES: A Cognitive-Affective Deep Model that Personalizes Math Problem Difficulty in Real Time and Improves Student Engagement with an Online Tutor
mmfp2880	Towards Adaptable Graph Representation Learning: An Adaptive Multi-Graph Contrastive Transformer
mmfp2282	Semantic-Aware Generator and Low-level Feature Augmentation for Few-shot Image Generation
mmfp2602	Language-guided Human Motion Synthesis with Atomic Actions
mmfp2740	MetaFBP: Learning to Learn High-Order Predictor for Personalized Facial Beauty Prediction
mmfp2939	PoSynDA: Multi-Hypothesis Pose Synthesis Domain Adaptation for Enhanced 3D Human Pose Estimation
mmfp2952	A Prior Instruction Representation Framework for Remote Sensing Image-text Retrieval
mmfp3029	Do Vision-Language Transformers Exhibit Visual Commonsense? An Empirical Study of VCR
mmfp3063	Real20M: A Large-scale E-commerce Dataset for Cross-domain Retrieval
mmfp3091	Multi-Granularity Interactive Transformer Hashing for Cross-modal Retrieval
mmfp3119	TeViS: Translating Text Synopses to Video Storyboards
mmfp3289	AdvCLIP: Downstream-agnostic Adversarial Examples in Multimodal Contrastive Learning
mmfp3609	UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons
mmfp3627	ATM: Action Temporality Modeling for Video Question Answering
mmfp4136	Pareto Invariant Representation Learning for Multimedia Recommendation
mmfp3241	Doubly Intention Learning for Cold-start Recommendation with Uncertainty-aware Stochastic Meta Process
mmfp3281	Learning Occlusion Disentanglement with Fine-grained Localization for Occluded Person Re-identification

All papers are presented also in the Poster Session T

Teaser Session T2

Paper id	Paper title
mmfp0152	Personalized Behavior-Aware Transformer for Multi-Behavior Sequential Recommendation
mmfp0189	SelfTalk: A Self-Supervised Commutative Training Diagram to Comprehend 3D Talking Faces
mmfp0409	Towards Accurate Lip-to-Speech Synthesis in-the-Wild
mmfp0570	Zero-Shot Learning by Harnessing Adversarial Samples
mmfp0660	Diffused Fourier Network for Video Action Segmentation
mmfp0682	Biased-Predicate Annotation Identification via Unbiased Visual Predicate Representation
mmfp0837	CCMB: A Large-scale Chinese Cross-modal Benchmark
mmfp0913	Chain-of-Look Prompting for Verb-centric Surgical Triplet Recognition in Endoscopic Videos
mmfp1026	Differentially Private Sparse Mapping for Privacy-Preserving Cross Domain Recommendation
mmfp1106	A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation
mmfp1209	Enhancing Multi-modal Multi-hop Question Answering via Structured Knowledge and Unified Retrieval-Generation
mmfp1309	Beyond Generic: Enhancing Image Captioning with Real-World Knowledge using Vision-Language Pre-Training Model
mmfp1312	Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels
mmfp1407	Layout-Guided High-Faithfulness Text-to-Image Generation
mmfp1456	Triple Correlations-Guided Label Supplementation for Unbiased Video Scene Graph Generation
mmfp1476	$\tt{DiVa}$: An Iterative Framework to Harvest More Diverse and Valid Labels from User Comments for Music
mmfp1487	Dark Knowledge Balance Learning for Unbiased Scene Graph Generation
mmfp1588	Revisiting Disentanglement and Fusion on Modality and Context in Conversational Multimodal Emotion Recognition
mmfp1757	Partial Annotation-based Video Moment Retrieval via Iterative Learning
mmfp1768	Online Distillation-enhanced Multi-modal Transformer for Sequential Recommendation
mmfp2055	AdaCLIP: Towards Pragmatic Multimodal Video Retrieval
mmfp2237	Knowledge Decomposition and Replay: A Novel Cross-modal Image-Text Retrieval Continual Learning Method
mmfp0209	Towards Explainable In-the-Wild Video Quality Assessment: a Database and a Language-Prompted Approach
mmfp0701	HSIC-based Moving Weight Averaging for Few-Shot Open-Set Object Detection
mmfp0711	Category-Level Articulated Object 9D Pose Estimation via Reinforcement Learning
mmfp1187	Rethinking the Localization in Weakly Supervised Object Localization
mmfp1377	Dual-Modal Attention-Enhanced Text-Video Retrieval with Triplet Partial Margin Contrastive Learning
mmfp1409	Generative Neutral Features-Disentangled Learning for Facial Expression Recognition
mmfp1509	Confidence-Aware Contrastive Learning for Semantic Segmentation
mmfp1523	Filling the Information Gap between Video and Query for Language-Driven Moment Retrieval
mmfp1686	Model Inversion Attack via Dynamic Memory Learning
mmfp1928	Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion
mmfp2066	Adversarial Training of Deep Neural Networks Guided by Texture and Structural Information
mmfp2089	TE-KWS: Text-Informed Speech Enhancement for Noise-Robust Keyword Spotting
mmfp0275	Sensing Micro-Motion Human Patterns using Multimodal mmRadar and Video Signal for Affective and Psychological Intelligence
mmfp2931	KeyPosS: Plug-and-Play Facial Landmark Detection through GPS-Inspired True-Range Multilateration

All papers are presented also in the Poster Session T