Offline Session T1

Paper id	Paper title
mmfp1868	Temporal Sentence Grounding in Streaming Videos
mmfp1973	QA-CLIMS: Question-Answer Cross Language Image Matching for Weakly Supervised Semantic Segmentation
mmfp2032	HSVLT: Hierarchical Scale-Aware Vision-Language Transformer for Multi-Label Image Classification
mmfp2083	GraphMedia:Communication-Balanced Graph Searching for Billion-scale Social Media Access
mmfp2120	Prototype-guided Knowledge Transfer for Federated Unsupervised Cross-modal Hashing
mmfp2138	Few-shot Multimodal Sentiment Analysis Based on Multimodal Probabilistic Fusion Prompts
mmfp2163	Progressive Visual Content Understanding Network for Image Emotion Classification
mmfp2196	FSNet: Frequency Domain Guided Superpixel Segmentation Network for Complex Scenes
mmfp2202	Topological Structure Learning for Weakly-Supervised Out-of-Distribution Detection
mmfp2233	Exploring Coarse-to-Fine Action Token Localization and Interaction for Fine-grained Video Action Recognition
mmfp2260	Towards End-to-End Unsupervised Saliency Detection with Self-Supervised Top-Down Context
mmfp2265	Whether you can locate or not? Interactive Referring Expression Generation
mmfp2291	Self-PT: Adaptive Self-Prompt Tuning for Low-Resource Visual Question Answering
mmfp2307	Transformer-based Point Cloud Generation Network
mmfp2360	GoRec: A Generative Cold-start Recommendation Framework
mmfp2390	Relational Contrastive Learning for Scene Text Recognition
mmfp2394	Expand BERT Representation with Visual Information via Grounded Language Learning with Multimodal Partial Alignment
mmfp2396	Prior Knowledge-driven Dynamic Scene Graph Generation with Causal Inference
mmfp2403	Knowledge Prompt-tuning for Sequential Recommendation
mmfp2411	Enhancing Sentence Representation with Visually-supervised Multimodal Pre-training
mmfp2415	Dynamic Triple Reweighting Network for Automatic Femoral Head Necrosis Diagnosis from Computed Tomography
mmfp2453	Scene Graph Masked Variational Autoencoders for 3D Scene Generation
mmfp2459	Visual Captioning at Will: Describing Images and Videos Guided by a Few Stylized Sentences
mmfp2503	MaTCR: Modality-Aligned Thought Chain Reasoning for Multimodal Task-Oriented Dialogue Generation
mmfp2516	Exploiting Low-confidence Pseudo-labels for Source-free Object Detection
mmfp2572	Text-based Person Search without Parallel Image-Text Data
mmfp2607	Rethinking Missing Modality Learning from a Decoding Perspective
mmfp2653	Handling Label Uncertainty for Camera Incremental Person Re-Identification
mmfp2662	ProtoHPE: Prototype-guided High-frequency Patch Enhancement for Visible-Infrared Person Re-identification
mmfp2714	Retrieval-based Knowledge Augmented Vision Language Pre-training
mmfp2717	Hierarchical Reasoning Network with Contrastive Learning for Few-Shot Human-Object Interaction Recognition
mmfp2791	Emotionally Situated Text-to-Speech Synthesis in User-Agent Conversation
mmfp2798	UniSA: Unified Generative Framework for Sentiment Analysis
mmfp2801	Enhancing Adversarial Robustness of Multi-modal Recommendation via Modality Balancing
mmfp2825	Video Entailment via Reaching a Structure-Aware Cross-modal Consensus
mmfp2832	MEDIC: A Multimodal Empathy Dataset in Counseling
mmfp2921	Contrastive Intra- and Inter-Modality Generation for Enhancing Incomplete Multimedia Recommendation
mmfp2935	MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion
mmfp2946	Learning Shared Semantic Information from Multimodal Bio-signals for Brain-Muscle Modulation Analysis
mmfp2962	SpaceCLIP: A Vision-Language Pretraining Framework With Spatial Reconstruction On Text
mmfp3005	Improving Zero-shot Visual Question Answering via Large Language Models with Reasoning Question Prompts
mmfp3013	CLIP-Hand3D: Exploiting 3D Hand Pose Estimation via Context-Aware Prompting
mmfp3016	Task-Adversarial Adaptation for Multi-modal Recommendation
mmfp3048	Mixup-Augmented Temporally Debiased Video Grounding with Multimodal Disentanglement
mmfp3076	Bridging Language and Geometric Primitives for Zero-shot Point Cloud Segmentation
mmfp3141	Interactive Interior Design Recommendation via Coarse-to-fine Multimodal Reinforcement Learning
mmfp3165	Multimodal Adaptive Emotion Transformer with Flexible Modality Inputs on A Novel Dataset with Continuous Labels
mmfp3226	Efficient Spatio-Temporal Video Grounding with Semantic-Guided Feature Decomposition
mmfp3334	Graph Spectral Perturbation for 3D Point Cloud Contrastive Learning
mmfp3474	Seeing in Flowing: Adapting CLIP for Action Recognition with Motion Prompts Learning
mmfp3480	Adaptive Contrastive Learning for Learning Robust Representations under Label Noise
mmfp3485	Style Transfer Meets Super-Resolution: Advancing Unpaired Infrared-to-Visible Image Translation with Detail Enhancement
mmfp3503	Counterfactual Cross-modality Reasoning for Weakly Supervised Video Moment Localization
mmfp3632	External Knowledge Dynamic Modeling for Image-text Retrieval
mmfp3653	Probability Distribution Based Frame-supervised Language-driven Action Localization
mmfp3690	Swin-UNIT: Transformer-based GAN for High-resolution Unpaired Image Translation
mmfp3816	Learning Semantics-Grounded Vocabulary Representation for Video-Text Retrieval

mmfp4000	Faster Attention with Heterogeneous RGCN for Medical ICD Coding Generation
mmfp4001	CMCU-CSS: Enhancing Naturalness via Commonsense-based Multi-modal Context Understanding in Conversational Speech Synthesis
mmfp4011	Modal-aware Bias Constrained Contrastive Learning for Multimodal Recommendation
mmfp4047	Prior-Guided Accuracy-Bias Tradeoff Learning for CTR Prediction in Multimedia Recommendation
mmfp4094	LUNA: Language as Continuing Anchors for Referring Expression Comprehension
mmfp4108	Null-text Guidance in Diffusion Models is Secretly a Cartoon-style Creator

mmfp1799	Your Negative May not Be True Negative: Boosting Image-Text Matching with False Negative Elimination
mmfp1865	Self-Distillation Dual-Memory Online Hashing with Hash Centers for Streaming Data Retrieval
mmfp1875	Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation
mmfp2004	RTQ: Rethinking Video-language Understanding Based on Image-text Model
mmfp2026	ChinaOpen: A Dataset for Open-world Multimodal Learning
mmfp2070	LocLoc: Low-level Cues and Local-area Guides for Weakly Supervised Object Localization
mmfp2095	Self-Supervised Cross-Language Scene Text Editing
mmfp2176	Dynamic Low-Rank Instance Adaptation for Universal Neural Image Compression
mmfp2194	Improving Cross-Modal Recipe Retrieval with Component-Aware Prompted CLIP Embedding
mmfp2206	Better Integrating Vision and Semantics for Improving Few-shot Classification
mmfp2246	MORE: A Multimodal Object-Entity Relation Extraction Dataset with a Benchmark Evaluation
mmfp2341	Face Encryption via Frequency-Restricted Identity-Agnostic Attacks
mmfp2380	DCEL: Deep Cross-Modal Evidential Learning for Text-Based Person Retrieval
mmfp2397	A Contrastive Learning Framework for Dual-Target Cross-Domain Recommendation
mmfp2551	Exploring Inconsistent Knowledge Distillation for Object Detection with Data Augmentation
mmfp2562	Hi-SIGIR: Hierachical Semantic-Guided Image-to-image Retrieval via Scene Graph
mmfp2683	Unlocking the Power of Multimodal Learning for Emotion Recognition in Conversation
mmfp2758	CONICA: A Contrastive Image Captioning Framework with Robust Similarity Learning
mmfp2760	Enhancing Visually-Rich Document Understanding via Layout Structure Modeling
mmfp2769	Dual Dynamic Proxy Hashing Network for Long-tailed Image Retrieval
mmfp2786	Spatio-Temporal Branching for Motion Prediction using Motion Increments
mmfp2819	Patch-Aware Representation Learning for Facial Expression Recognition
mmfp2833	Towards Visual Taxonomy Expansion
mmfp2843	Joint Searching and Grounding: Multi-Granularity Video Content Retrieval
mmfp2870	Modality-agnostic Augmented Multi-Collaboration Representation for Semi-supervised Heterogenous Face Recognition
mmfp2873	Reducing Intrinsic and Extrinsic Data Biases for Moment Localization with Natural Language
mmfp2880	Towards Adaptable Graph Representation Learning: An Adaptive Multi-Graph Contrastive Transformer
mmfp2927	Precise Target-Oriented Attack against Deep Hashing-based Retrieval
mmfp2930	MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition
mmfp2960	Underwater Image Enhancement by Transformer-based Diffusion Model with Non-uniform Sampling for Skip Strategy
mmfp2990	Dynamic Contrastive Learning with Pseudo-samples Intervention for Weakly Supervised Joint Video MR and HD
mmfp3054	Fine-Grained Visual Prompt Learning of Vision-Language Models for Image Recognition
mmfp3084	ACQ: Few-shot Backdoor Defense via Activation Clipping and Quantizing
mmfp3172	Iterative Learning with Extra and Inner Knowledge for Long-tail Dynamic Scene Graph Generation
mmfp3300	PointCRT: Detecting Backdoor in 3D Point Cloud via Corruption Robustness
mmfp3301	When Perceptual Authentication Hashing Meets Neural Architecture Search
mmfp3303	Local Consensus Enhanced Siamese Network with Reciprocal Loss for Two-view Correspondence Learning
mmfp3313	Slow-Fast Time Parameter Aggregation Network for Class-Incremental Lip Reading
mmfp3341	Towards Deconfounded Image-text Matching with Causal Inference
mmfp3624	Progressive Positive Association Framework for Image and Text Retrieval
mmfp3661	Contrast-augmented Diffusion Model with Fine-grained Sequence Alignment for Markup-to-Image Generation
mmfp3735	Zero-Shot Object Detection by Semantics-Aware DETR with Adaptive Contrastive Loss
mmfp3815	Deconfounded Visual Question Generation with Causal Inference
mmfp3872	Focusing on Flexible Masks: A Novel Framework for Panoptic Scene Graph Generation with Relation Constraints
mmfp3907	WormTrack: Dataset and Benchmark for Multi-Object Tracking in Worm Crowds
mmfp3921	Category-Specific Prompts for Animal Action Recognition with Pretrained Vision-Language Models
mmfp3927	Prototype-guided Cross-modal Completion and Alignment for Incomplete Text-based Person Re-identification
mmfp3935	Multimodal Physiological Signals Fusion for Online Emotion Recognition
mmfp4038	Zero-TextCap: Zero-shot Framework for Text-based Image Captioning
mmfp4120	A Multitask Framework for Graffiti-to-Image Translation
mmfp4139	Scene-text Oriented Visual Entailment: Task, Dataset and Solution
mmfp4170	Unsupervised Hashing with Contrastive Learning by Exploiting Similarity Knowledge and Hidden Structure of Data

Offline Session T2

Paper id	Paper title
mmfp0012	Generalizable Label Distribution Learning
mmfp0052	Shift Pruning: Equivalent Weight Pruning for CNN via Differentiable Shift Operator
mmfp0054	Learning from More: Combating Uncertainty Cross-multidomain for Facial Expression Recognition
mmfp0065	Resource Constrained Model Compression via Minimax Optimization for Spiking Neural Networks
mmfp0068	Emo-DNA: Emotion Decoupling and Alignment Learning for Cross-Corpus Speech Emotion Recognition
mmfp0232	NightHazeFormer: Single Nighttime Haze Removal Using Prior Query Transformer
mmfp0264	V2Depth: Monocular Depth Estimation via Feature-Level Virtual-View Simulation and Refinement
mmfp0295	Reservoir Computing Transformer for Image-Text Retrieval
mmfp0309	Invariant Meets Specific: A Scalable Harmful Memes Detection Framework
mmfp0327	Enhancing Domain-Invariant Parts for Generalized Zero-Shot Learning
mmfp0351	Symmetrical Linguistic Feature Distillation with CLIP for Scene Text Recognition
mmfp0410	Avatar Knowledge Distillation: Self-ensemble Teacher Paradigm with Uncertainty
mmfp0421	Fine-grained Pseudo Labels for Scene Text Recognition
mmfp0441	Striking a Balance: Unsupervised Cross-Domain Crowd Counting via Knowledge Diffusion
mmfp0472	All in One: Exploring Unified Vision-Language Tracking with Multi-Modal Alignment
mmfp0484	Multimodal Prompt Transformer with Hybrid Contrastive Learning for Emotion Recognition in Conversation
mmfp0541	Vision-Guided Composed Image Retrieval
mmfp0548	Adaptive Decoupled Pose Knowledge Distillation
mmfp0564	Semi-Supervised Convolutional Vision Transformer with Bi-Level Uncertainty Estimation for Medical Image Segmentation
mmfp0583	COPA : Efficient Vision-Language Pre-training through Collaborative Object- and Patch-Text Alignment
mmfp0595	RAMM: Retrieval-augmented Biomedical Visual Question Answering with Multi-modal Pre-training
mmfp0628	VPA: Fully Test-Time Visual Prompt Adaptation
mmfp0661	RetouchingFFHQ: A Large-scale Dataset for Fine-grained Face Retouching Detection
mmfp0769	SUR-adapter: Enhancing Text-to-Image Pre-trained Diffusion Models with Large Language Models
mmfp0776	Stepwise Refinement Short Hashing for Image Retrieval
mmfp0779	Multi-View Graph Convolutional Network for Multimedia Recommendation
mmfp0865	Conversational Composed Retrieval with Iterative Sequence Refinement
mmfp0887	Zero-shot Skeleton-based Action Recognition via Mutual Information Estimation and Maximization
mmfp0902	CPLFormer: Cross-scale Prototype Learning Transformer for Image Snow Removal
mmfp0923	Dense Object Grounding in 3D Scenes
mmfp0938	Semi-supervised Semantic Segmentation with Mutual Knowledge Distillation
mmfp1019	Light-VQA: A Multi-Dimensional Quality Assessment Model for Low-Light Video Enhancement
mmfp1028	Non-Exemplar Class-Incremental Learning via Adaptive Old Class Reconstruction
mmfp1045	Semi-supervised Deep Multi-view Stereo
mmfp1091	Relation Triplet Construction for Cross-modal Text-to-Video Retrieval
mmfp1107	DAE-Talker: High Fidelity Speech-Driven Talking Face Generation with Diffusion Autoencoder
mmfp1131	Learning Style-Invariant Robust Representation for Generalizable Visual Instance Retrieval
mmfp1147	A Method of Micro-Geometric Details Preserving in Surface Reconstruction from Gradient
mmfp1218	Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation
mmfp1232	AesCLIP: Multi-Attribute Contrastive Learning for Image Aesthetics Assessment
mmfp1277	Hashing One With All
mmfp1284	Semantics-Enriched Cross-Modal Alignment for Complex-Query Video Moment Retrieval
mmfp1364	Toward High Quality Facial Representation Learning
mmfp1374	Uncertainty-Driven Dynamic Degradation Perceiving and Background Modeling for Efficient Single Image Desnowing
mmfp1386	Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval
mmfp1421	Hyperspectral Image Denoising with Spectrum Alignment
mmfp1444	Sparse Sharing Relation Network for Panoptic Driving Perception
mmfp1447	Capturing Co-existing Distortions in User-Generated Content for No-reference Video Quality Assessment
mmfp1547	Attributes Grouping and Mining Hashing for Fine-Grained Image Retrieval
mmfp1583	General Debiasing for Multimodal Sentiment Analysis
mmfp1678	PixelFace+: Towards Controllable Face Generation and Manipulation with Text Descriptions and Segmentation Masks
mmfp1681	Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer
mmfp1770	Isolation and Induction: Training Robust Deep Neural Networks against Model Stealing Attacks
mmfp1780	Learning Implicit Entity-object Relations by Bidirectional Generative Alignment for Multimodal NER
mmfp1784	Constructing Holistic Spatio-Temporal Scene Graph for Video Semantic Role Labeling
mmfp1789	HAAN: Human Action Aware Network for Multi-label Temporal Action Detection
mmfp1803	Giving text more imagination space for image-text matching

mmfp1811	Ground-to-Aerial Person Search: Benchmark Dataset and Approach
mmfp1828	GCMA: Generative Cross-Modal Transferable Adversarial Attacks from Images to Videos
mmfp1847	Sketch Input Method Editor: A Comprehensive Dataset and Methodology for Systematic Input Recognition
mmfp1857	CARIS: Context-Aware Referring Image Segmentation

mmfp0040	Orthogonal Uncertainty Representation of Data Manifold for Robust Long-Tailed Learning
mmfp0066	Unlocking the Power of Cross-Dimensional Semantic Dependency for Image-Text Matching
mmfp0075	VioLET: Vision-Language Efficient Tuning with Collaborative Multi-modal Gradients
mmfp0088	Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark
mmfp0101	Depth-Aware Sparse Transformer for Video-Language Learning
mmfp0129	Cerebrovascular Segmentation in TOF-MRA with Topology Regularization Adversarial Model
mmfp0135	POAR: Towards Open-World Pedestrian Attribute Recognition
mmfp0156	Emotion-Prior Awareness Network for Emotional Video Captioning
mmfp0174	Enhancing Real-Time Super Resolution with Partial Convolution and Efficient Variance Attention
mmfp0192	Improving Human-Object Interaction Detection via Virtual Image Learning
mmfp0206	Mind the Gap: Improving Success Rate of Vision-and-Language Navigation by Revisiting Oracle Success Routes
mmfp0221	Zero-shot Micro-video Classification with Neural Variational Inference in Graph Prototype Network
mmfp0228	Sequential Affinity Learning for Video Restoration
mmfp0235	AdaBrowse: Adaptive Video Browser for Efficient Continuous Sign Language Recognition
mmfp0270	Separate and Locate: Rethink the Text in Text-based Visual Question Answering
mmfp0349	Beat: Bi-directional One-to-Many Embedding Alignment for Text-based Person Retrieval
mmfp0412	CLIP-Count: Towards Text-Guided Zero-Shot Object Counting
mmfp0426	DAOT: Domain-Agnostically Aligned Optimal Transport for Domain-Adaptive Crowd Counting
mmfp0448	Feeling Positive? Predicting Emotional Image Similarity from Brain Signals
mmfp0486	Hybrid Interaction Temporal Knowledge Graph Embedding Based on Householder Transformations
mmfp0505	ZRIGF: An Innovative Multimodal Framework for Zero-Resource Image-Grounded Dialogue Generation
mmfp0534	Improving Semi-Supervised Semantic Segmentation with Dual-Level Siamese Structure Network
mmfp0562	Towards Real-Time Sign Language Recognition and Translation on Edge Devices
mmfp0601	Learning Comprehensive Representations with Richer Self for Text-to-Image Person Re-Identification
mmfp0722	Cross-modal Contrastive Learning for Multimodal Fake News Detection
mmfp0729	Mirror-NeRF: Learning Neural Radiance Fields for Mirrors with Whitted-Style Ray Tracing
mmfp0732	LiFT: Transfer Learning in Vision-Language Models for Downstream Adaptation and Generalization
mmfp0758	StableVQA: A Deep No-Reference Quality Assessment Model for Video Stability
mmfp0789	Auditory Attention Decoding with Task-Related Multi-View Contrastive Learning
mmfp0840	A Reinforcement Learning-Based Automatic Video Editing Method Using Pre-trained Vision-Language Model
mmfp0844	Unsupervised Domain Adaptation for Referring Semantic Segmentation
mmfp0847	EAT: An Enhancer for Aesthetics-Oriented Transformers
mmfp0867	Semantic-Guided Feature Distillation for Multimodal Recommendation
mmfp0914	Prompt Me Up: Unleashing the Power of Alignments for Multimodal Entity and Relation Extraction
mmfp0934	Uniformly Distributed Category Prototype-Guided Vision-Language Framework for Long-Tail Recognition
mmfp0935	Language-Guided Visual Aggregation Network for Video Question Answering
mmfp0960	Distilling Vision-Language Foundation Models: A Data-Free Approach via Prompt Diversification
mmfp0983	Blind Image Super-resolution with Rich Texture-Aware Codebook
mmfp0994	C^{2}MR: Continual Cross-Modal Retrieval for Streaming Multi-modal Data
mmfp1033	Spatial-angular Quality-aware Representation Learning for Blind Light Field Image Quality Assessment
mmfp1049	Fine-Grained Spatiotemporal Motion Alignment for Contrastive Video Representation Learning
mmfp1099	Feature-Suppressed Contrast for Self-Supervised Food Pre-training
mmfp1317	Semi-supervised Domain Adaptation via Joint Contrastive Learning with Sensitivity
mmfp1347	Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models
mmfp1353	BMI-Net: A Brain-inspired Multimodal Interaction Network for Image Aesthetic Assessment
mmfp1354	Deep Algorithm Unrolling with Registration Embedding for Pansharpening
mmfp1439	Weakly-supervised Video Scene Graph Generation via Unbiased Cross-modal Learning
mmfp1445	AvatarFusion: Zero-shot Generation of Clothing-Decoupled 3D Avatars Using 2D Diffusion
mmfp1458	Distortion-aware Transformer in 360° Salient Object Detection
mmfp1652	OCSKB: An Object Component Sketch Knowledge Base for Fast 6D Pose Estimation
mmfp1672	Enhancing Product Representation with Multi-form Interactions for Multimodal Conversational Recommendation
mmfp1691	Graph to Grid: Learning Deep Representations for Multimodal Emotion Recognition
mmfp1712	VCMaster: Generating Diverse and Fluent Live Video Comments Based on Multimodal Contexts
mmfp1717	Facial Auto Rigging from 4D Expressions via Skinning Decomposition