Introduction to the Special Issue on Deep Multimodal Generation and Retrieval

Fei, Hao; Ji, Wei; Wei, Yinwei; Zheng, Zhedong; Shen, Jialie; Hanjalic, Alan; Zimmermann, Roger

DOI
10.1145/3762666

Publication date
2025

Document Version
Final published version

Published in
ACM Transactions on Multimedia Computing, Communications and Applications

Citation (APA)
Fei, H., Ji, W., Wei, Y., Zheng, Z., Shen, J., Hanjalic, A., & Zimmermann, R. (2025). Introduction to the Special Issue on Deep Multimodal Generation and Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, *21*(11), Article 304. https://doi.org/10.1145/3762666

Published: 10 November 2025
Online AM: 05 September 2025
Accepted: 14 August 2025
Revised: 13 August 2025
Received: 13 August 2025

INTRODUCTION

Introduction to the Special Issue on Deep Multimodal Generation and Retrieval

HAO FEI, National University of Singapore, Singapore City, Singapore

WEI JI, National University of Singapore, Singapore City, Singapore

YINWEI WEI, Monash University, Melbourne, VIC, Australia

ZHEDONG ZHENG, University of Macau, Taipa, Macao

JIALIE SHEN, St George's, University of London, London, U.K.

ALAN HANJALIC, Delft University of Technology, Delft, Zuid-Holland, Netherlands

View all

Introduction to the Special Issue on Deep Multimodal Generation and Retrieval

1 Introduction

The rapid development of artificial intelligence has significantly advanced how machines acquire and interact with multimodal information [8, 10, 30, 37, 43, 67, 68]. Information acquisition, once confined to unimodal text-based pipelines, has now expanded to embrace **multimodal generation and retrieval (MMGR)** systems that jointly leverage visual, linguistic, and structured modalities [7, 18–21, 35, 38, 46, 50, 59, 62, 69, 74]. These systems are becoming increasingly crucial in various real-world applications, such as content creation, cross-modal search, digital assistants, and scientific knowledge discovery [3, 5, 36].

Traditionally, the fields of **information generation (IG)** and **information retrieval (IR)** have focused on isolated unimodal settings, largely limited to text-based content and indexing [2, 4, 11, 12, 22, 28, 29, 33, 53, 54, 57, 60, 72]. Computer graphics and vision researchers, on the other hand, have primarily focused on visual content creation and manipulation, such as generating and processing images and videos [15–17, 31, 42, 63, 65]. However, real-world information is inherently multimodal, requiring systems to reason jointly over images, videos, language, audio, and structured signals [9, 13, 26, 47, 48, 51, 52, 55, 70]. This has led to the rise of MMGR research [23, 58, 61, 71, 73], which calls for effective fusion, alignment, and interpretation of heterogeneous modalities in both forward generation and backward retrieval processes. Recent advances in large-scale pretraining and generative models have further enabled scalable MMGR systems, bridging vision-language gaps and enabling controllable synthesis [1, 27, 49, 76].

Nevertheless, key challenges remain unresolved in the MMGR domain. These challenges span algorithmic, systemic, and robustness aspects, including:

— *Learning generalized and transferable multimodal representations* across domains and modalities is crucial for robustness. Recent works explore spectrum recombination [24], unsupervised visual-language pretraining [27], and triplet contrastive representation learning [39] to enhance representation transferability.

CCS Concepts: • **Information systems** → **Information retrieval**; **Computing methodologies** → *Knowledge representation and reasoning*; **Computer vision**;

Additional Key Words and Phrases: Gesture recognition, pose estimation, sign language translation, generative models, adversarial learning, behavior analysis, explainability, accessibility

ACM Reference format:
Hao Fei, Wei Ji, Yinwei Wei, Zhedong Zheng, Jialie Shen, Alan Hanjalic, and Roger Zimmermann. 2025. Introduction to the Special Issue on Deep Multimodal Generation and Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 11, Article 304 (November 2025), 13 pages.
https://doi.org/10.1145/3762666

ACM 1551-6865/2025/11-ART304
https://doi.org/10.1145/3762666

— *Ensuring semantic alignment in generation and retrieval* remains difficult, especially under sparse supervision. Efforts such as spatial-temporal attention for video-text retrieval [40], object-pose disentanglement for editing [66], and correlation-aware attention in fashion matching [5] address these issues with refined architectures.
— *Generating controllable and coherent multimodal content* is critical for applications like text-to-image, sketch-based editing, and 3D avatar generation. Recent advances include prompt-sensitive diffusion modeling [1], sketch-guided GANs [41], and 3D avatar generation using face priors [14].
— *Improving interpretability and reducing hallucination* is vital for responsible AI. Works such as multimodal consistency suppression for fake news [44], emotion-cause extraction with holistic constraints [25], and zero-shot sarcasm detection with **vision-language large models (VLLMs)** [45] highlight emerging approaches to enhance explainability.
— *Adapting large models efficiently to new domains* requires methods like in-context tuning [3], efficient distillation for small object editing [56], and prompt learning without supervision [49], which aim to reduce training cost and data dependence.
— *Designing unified evaluation protocols and real-world benchmarks* is essential for scalable deployment. Studies address this through protein-language generation tasks [64], few-shot segmentation with multimodal large models [75], and visible-infrared person ReID with label refinement [6].

This Special Issue on *Deep MMGR* presents rigorously reviewed papers that collectively tackle the above challenges from novel angles. Contributions span across text-video retrieval [40], 3D content generation [34], noise-robust retrieval [32], sociocultural dialogue modeling [36], and bias mitigation [32]. Together, they reflect the state-of-the-art in MMGR and illuminate new directions for building generalizable, controllable, and trustworthy multimodal AI systems.

2 Overview of the Publications Included in This Special Issue

This special issue compiles twenty-one high-quality papers, which collectively address critical challenges in deep MMGR. These contributions span four thematic categories: (1) Multimodal Semantics Understanding, (2) Generative Models for Vision Synthesis, (3) **Multimodal Information Retrieval (MMIR)**, and (4) Explainable and Reliable Multimodal Learning. Each paper proposes new methods to improve cross-modal alignment, generative quality, retrieval effectiveness, and interpretability in multimodal learning systems. Table 1 provides a structured summary of the contributions.

2.1 Multimodal Semantics Understanding

Understanding and modeling semantics across modalities is central to multimodal intelligence. The papers in this category address core challenges including alignment of visual and linguistic representations, cross-modal reasoning, semantic parsing, robustness to noise and weak supervision, and efficient tuning for downstream tasks. These contributions significantly advance the understanding of multimodal semantics through pretraining strategies, emotion-cause analysis, sarcasm detection, and label-efficient learning.

Contrastive Visual-Language Pretraining without Annotations. Li et al. [27] propose CVLP-NaVD, a contrastive visual-language pretraining framework tailored for *non-annotated visual description* generation. This work tackles the challenge of learning grounded multimodal representations without paired text, leveraging noise-aware visual prompts and synthetic captions produced by CLIP and

Table 1. Overview of Articles: Categories, Contributions, and Addressed Challenges

Category	Paper	Main Contribution	Challenges Addressed	Generalization	Semantic Align	Faithful Gen.	Explainability	Efficient Tuning	Benchmarking
Multimodal Semantics Understanding	Li et al. [27]	Contrastive pretraining for visual-language without annotations	Transfer, weakly supervised learning	✓	✓			✓	
	Zhang et al. [64]	Protein sequence-to-natural language captioning	Scientific alignment across modalities	✓	✓				
	Li et al. [25]	Emotion-cause pair extraction with label constraints	Semantic interaction modeling		✓		✓		
	Chen et al. [3]	MMICT	Prompt-based fine-tuning framework					✓	
	Zhou et al. [75]	Few-shot segmentation with multimodal LLMs	Orthogonal space learning	✓				✓	
	Liu et al. [32]	Bias mitigation and retrieval optimization	Cross-modal robustness	✓	✓		✓		
	Wang et al. [45]	VLLM-based zero-shot multimodal sarcasm detection	Zero-shot, semantic ambiguity		✓		✓		
Generative Models for Vision Synthesis	Gan et al. [14]	3D avatar generation from unseen expressions	High-fidelity generation			✓			
	Cai et al. [1]	Prompt analysis for text-to-image diffusion	Temporal-spatial generative insights		✓	✓	✓		
	Wei et al. [49]	Domain-generalized image captioning via prompt learning	Robust caption transfer	✓		✓			
	Zhang et al. [66]	Pose-object controlled image editing using SAM	Fine-grained image control		✓	✓		✓	
	Sun et al. [41]	Sketch-based fashion GAN with content decoupling	Style disentanglement			✓			
	Ma et al. [34]	Bridging text-to-2D and 3D content generation	Cross-modal 3D generation	✓	✓	✓			
	Wu et al. [56]	Distillation for small object editing	Efficient editing pipeline			✓		✓	
MMIR	Shen et al. [39]	Triplet contrastive learning for vehicle ReID	Unsupervised metric alignment	✓				✓	
	Cui et al. [5]	Cross-modal attention for fashion matching	Visual-semantic outfit retrieval		✓				
	Shen et al. [40]	Spatio-temporal attention for video-text retrieval	Video-language alignment	✓	✓				
	Dai et al. [6]	Label refinement for visible-infrared ReID	Dual-modal optimization					✓	
	Jing et al. [24]	Transformer for hyperspectral classification	Modality-specific spectral reasoning	✓				✓	
Explainable and Reliable Multimodal Learning	Qu et al. [36]	Frame-based sociocultural norm base construction	Dialogue social grounding				✓		✓
	Tao et al. [44]	Fake news detection with multimodal suppression factor	Consistency suppression, reliability		✓		✓		✓

MMICT, Multimodal in-context tuning.

BLIP2. Through multi-instance contrastive alignment and dual prompt tuning, the model demonstrates strong performance on image captioning tasks across domains. This approach aligns with the theme of multimodal semantics by enabling *structure-aware alignment* and *robust representation learning* under weak supervision settings.

Protein Captioning across Scientific Modalities. Zhang et al. [64] present a pioneering effort to bridge the gap between protein sequences and natural language via protein captioning. This work explores a novel scientific modality where the input is a protein sequence, and the output is a human-readable textual description. The authors design a Transformer-based captioning architecture and introduce tokenization techniques suitable for biological data. The results show promising accuracy and coherence, opening new possibilities in biomolecular interpretation. The work expands multimodal semantics into the scientific domain, tackling challenges in *semantic alignment across heterogeneous modalities* and demonstrating *efficiency* in low-resource biological applications.

Emotion-Cause Pair Extraction with Interaction-Aware Constraints. Li et al. [25] develop a holistic interaction framework for **multimodal emotion-cause pair extraction (MEPE)** that integrates cross-modal interaction modeling and label constraints. The authors propose a fusion module combining textual and visual features, enriched with hierarchical dependencies and global-to-local pair matching. The model is trained with a multi-objective loss that balances emotion recognition, cause detection, and their interaction. Evaluated on standard MEPE benchmarks, the framework outperforms previous baselines, particularly in cases involving subtle cause triggers. This work directly contributes to semantic-level understanding by enhancing *fine-grained reasoning between affective cues and their triggers.*

Multimodal In-Context Tuning (MMICT). Chen et al. [3] propose MMICT, a novel method for *multimodal fine-tuning via in-context examples*. Inspired by prompt-based learning in language models, the framework introduces a hybrid encoder for image-text tasks that generates pseudo in-context demonstrations during training. These prompts improve generalization by injecting task-aware conditioning, enabling efficient few-shot adaptation. The study validates MMICT across multiple downstream tasks, including VQA and image captioning, demonstrating superior performance with minimal supervision. This work is critical for scalable and efficient multimodal semantic understanding, especially in low-resource scenarios.

Orthogonal Space Learning with Multimodal Large Models. Zhou et al. [75] explore the use of orthogonal semantic subspaces to improve few-shot learning in multimodal segmentation. The authors introduce a framework that separates modality-specific and shared representations in a latent orthogonal space and guides few-shot learning through self-adaptive prototype construction. Using large pretrained vision-language models as the backbone, this approach improves generalization on unseen segmentation classes. This article directly tackles the challenge of *scalability and generalization* in multimodal understanding, with a method grounded in representation disentanglement and task transfer.

Bias-Robust Cross-Modal Retrieval. Liu et al. [32] address the issue of *bias and representation collapse* in cross-modal retrieval by introducing an end-to-end framework that integrates bias mitigation with robust representation learning. A modality-sensitive attention mechanism coupled with adversarial disentanglement reduces inter-modal noise and improves generalization. Experimental results on benchmark datasets show significant improvements in retrieval accuracy under domain-shift settings. This work enhances multimodal semantic grounding by promoting *robustness to distributional bias and noisy modality-specific signals.*

Zero-Shot Sarcasm Detection via VLLM Agent. Wang et al. [45] propose S^3 Agent, a versatile framework for *zero-shot sarcasm detection* using VLLMs. The agent reformulates sarcasm recognition as a structured understanding problem, leveraging multimodal prompts and CoT reasoning. By integrating perception and reflection modules, S^3 Agent captures nuanced visual-textual incongruity that defines sarcasm. Evaluated on multimodal sarcasm datasets, the model achieves state-of-the-art performance without fine-tuning. This article exemplifies how *semantic abstraction and alignment* from large multimodal models can be harnessed to address highly subjective and subtle semantic phenomena.

2.2 Generative Models for Vision Synthesis

Generative modeling plays a vital role in multimodal AI, enabling controllable content synthesis across image, video, and 3D domains. This section presents seven papers that tackle core challenges in vision synthesis, such as generalization across domains, faithful alignment with conditioning inputs (e.g., prompts or sketches), and efficiency under resource constraints.

ExpAvatar: 3D Prior-Based Expressive Avatar Generation. In "*ExpAvatar: High-Fidelity Avatar Generation of Unseen Expressions with 3D Face Priors*," Gan et al. [14] propose a novel framework for generating expressive avatars by leveraging 3D face priors to disentangle identity and expression representations. The method introduces a dynamic blendshape reconstruction module that encodes prior facial geometry, followed by a high-resolution texture generation network to synthesize expressive avatars. By explicitly modeling unseen expressions using a latent fusion strategy, ExpAvatar effectively generalizes to expressions not seen during training and achieves high-fidelity synthesis results. This work directly addresses challenges of semantic alignment and faithful generation in expression-driven avatar creation.

Picasso: Prompt Design Analysis in Diffusion Models. Cai et al. [1] present "*Picasso: Analyzing Prompt Design for Text-to-Image Generative Diffusion Models from a Temporal-Spatial Perspective*," a systematic investigation into how prompts influence image generation quality in diffusion models. The authors build a benchmark that quantifies prompt controllability across time (generation steps) and space (layout or region effects), revealing significant inconsistencies in model interpretability. They propose a temporal-spatial consistency score and demonstrate through experiments that prompt engineering has implicit biases across diffusion iterations. This work contributes to the challenge of improving explainability and controllability in large-scale generative models.

Unsupervised Prompt Learning for Image Captioning. Wei and Chen [49] propose "*Improving Domain Generalization for Image Captioning with Unsupervised Prompt Learning*," which aims to enhance captioning models across diverse domains without requiring domain-specific annotations. They design an unsupervised prompt generator based on self-supervised contrastive learning and integrate it into a captioning model to guide semantic focus. Their experiments on multiple out-of-domain datasets (e.g., nocaps, flickr30k) show that the prompt-enhanced captioner outperforms standard models under distribution shifts, showcasing improvements in both fluency and relevance. The work addresses scalability and domain generalization in caption generation.

SAMControl: Pose and Object Controllable Image Editing. The paper "*SAMControl: Controlling Pose and Object for Image Editing with Soft Attention Mask*," by Zhang et al. [66], introduces a soft-attention mask mechanism that enables users to control both pose and object appearance during image editing. Built atop a U-Net-based generator and guided by textual inputs and object masks, the model disentangles foreground transformations from background consistency. The

system achieves flexible editing with minimal manual tuning and supports real-time interactions. It directly tackles controllable content generation, semantic alignment, and efficient inference.

CoDE-GAN: Sketch-Guided Fashion Editing. In "*CoDE-GAN: Content Decoupled and Enhanced GAN for Sketch-Guided Flexible Fashion Editing*," Sun et al. [41] address the challenge of aligning sparse sketch inputs with rich appearance representations. The proposed GAN architecture decouples content and style using dual encoders and integrates a spatial attention mechanism to fuse sketch structure with clothing features. The model allows flexible attribute editing (e.g., texture, sleeve, collar) and achieves realistic generation on fashion datasets such as DeepFashion. The method enhances editing fidelity and visual realism in style transfer scenarios.

3D Content Synthesis by Bridging Text-to-2D and Text-to-3D. Ma et al. [34] introduce "*Creating High-Quality 3D Content by Bridging the Gap Between Text-to-2D and Text-to-3D Generation*," proposing a two-stage framework that first generates high-quality 2D images from text, then lifts them into 3D assets using neural surface reconstruction. The novelty lies in transferring learned priors from 2D diffusion models into the 3D generation space without explicit 3D supervision. Extensive experiments show that their method surpasses traditional text-to-3D pipelines in geometric fidelity and semantic consistency. This study tackles multimodal fusion and efficiency in 3D content creation.

SOEDiff: Efficient Small Object Editing via Distillation. Finally, Wu et al. [56] present "*SOEDiff: Efficient Distillation for Small Object Editing*," a lightweight model designed for precise editing of small-scale visual regions. The authors propose a region-aware distillation strategy that preserves context while modifying target objects, supported by attention-guided reconstruction and diffusion-based refinement. Their experiments on object editing benchmarks demonstrate substantial improvements in editing precision with reduced computational overhead. SOEDiff addresses the challenges of efficient generation and region-level fidelity.

2.3 MMIR

MMIR targets the alignment, fusion, and matching across heterogeneous data sources such as images, videos, and textual descriptions. The selected works in this section address critical challenges in MMIR, including label scarcity, cross-modal fusion, temporal modeling, and modality-specific reasoning. They explore innovative methods ranging from contrastive representation learning to dual-modality refinement, pushing the boundaries of efficient and robust retrieval across visual and textual modalities.

Triplet Contrastive Learning for Unsupervised Vehicle Re-Identification. Shen et al. [39] introduce "*Triplet Contrastive Representation Learning for Unsupervised Vehicle Re-Identification*," a method that alleviates the reliance on labeled data by introducing a novel framework for unsupervised representation learning. The model builds upon a hierarchical clustering approach to generate pseudo-labels, which are refined through a momentum-updated memory bank and hard sample mining. A triplet-based contrastive loss is then applied to learn robust vehicle embeddings under varying lighting and pose conditions. Experimental results on VeRi-776 and VehicleID datasets show substantial improvements over existing unsupervised baselines, directly addressing the challenges of unsupervised retrieval and cross-domain robustness in vehicle re-ID.

Cross-Modal Attention for Outfit Compatibility Modeling. Cui et al. [5] present "*Correlation-Aware Cross-Modal Attention Network for Fashion Compatibility Modeling in UGC Systems*," which models the compatibility among fashion items using multimodal cues from both text and image. The

method introduces a correlation-aware cross-attention module that dynamically adjusts attention weights based on semantic alignment between product descriptions and visual appearances. This fine-grained fusion allows the model to reason over **user-generated content (UGC)** and enhances compatibility prediction. Experiments on large-scale fashion datasets demonstrate notable gains in recommendation precision and interpretability, addressing semantic fusion and alignment challenges in UGC-based multimodal systems.

Spatio-Temporal Attention for Video-Text Retrieval. In "*Spatio-Temporal Attention for Text-Video Retrieval*," Shen et al. [40] propose a dual-stream architecture that models fine-grained temporal and spatial correspondences between video clips and textual queries. The system leverages a spatio-temporal attention fusion module to capture the evolving semantic relevance across video frames and sentence elements. A contrastive loss aligns matched video-text pairs while maintaining intra-modal consistency. The approach achieves competitive retrieval accuracy on standard benchmarks such as MSR-VTT and LSMDC, advancing temporal grounding and fusion quality in cross-modal video understanding.

Dual-Modality-Shared Learning for Infrared Person ReID. Dai et al. [6] explore "*Dual-Modality-Shared Learning and Label Refinement for Unsupervised Visible-Infrared Person Re-Identification*," focusing on the challenging task of matching person identities across visible and infrared modalities. The proposed framework includes two key modules: a modality-shared representation learning unit that minimizes cross-modality discrepancy and a label refinement mechanism that iteratively updates pseudo-labels through confidence re-estimation. The model shows superior generalization performance on SYSU-MM01 and RegDB datasets. This work tackles cross-spectral matching and unsupervised learning, enabling more reliable person retrieval in surveillance applications.

SpectrumRecombineFormer for Hyperspectral Image Classification. Finally, Jing et al. [24] present "*SRF: SpectrumRecombineFormer for Hyperspectral Image Classification*," a transformer-based framework that introduces a spectrum recombination module to exploit modality-specific spectral bands effectively. By adaptively recombining bands into structurally meaningful subspaces and applying self-attention, SRF enhances feature representation quality while preserving class-discriminative information. Experiments on public hyperspectral datasets such as Indian Pines and Pavia University validate the method's superiority over CNN and standard transformer baselines. The study addresses modality-specific reasoning and scalable modeling for hyperspectral visual retrieval.

2.4 Explainable and Reliable Multimodal Learning

As AI systems increasingly interact with human users in socially consequential scenarios, the demand for explainability, reliability, and societal alignment in multimodal learning becomes paramount. The selected works in this section address these imperatives by incorporating structured commonsense, sociocultural grounding, and consistency-aware modeling into multimodal architectures. These studies explore how to model the trustworthiness and interpretability of multimodal reasoning systems, especially in the contexts of dialogue and misinformation detection.

Frame-Based Sociocultural Norm Induction for Dialogue Agents. Qu et al. [36] propose "*Scalable Frame-Based Construction of Sociocultural Norm Bases for Socially Aware Dialogues*," which seeks to integrate sociocultural awareness into dialogue systems through an explainable, structured knowledge representation. The authors build upon the concept of frame semantics by constructing a large-scale sociocultural norm base derived from commonsense knowledge, behavioral datasets, and large-scale pretraining corpora. A key contribution is the formulation of a multi-stage norm

induction pipeline that identifies action-event tuples and links them to culturally informed expectations via structured templates. This enriched knowledge base is then integrated into dialogue agents to support norm-aware response generation and violation detection. By grounding outputs in explicit sociocultural frames, the model promotes transparency and interpretability. Empirical results demonstrate improved dialogue quality and norm adherence across several human-in-the-loop evaluation settings, thus directly addressing the challenges of societal alignment, symbolic reasoning, and trust in language-based multimodal AI.

Consistency Suppression for Multimodal Fake News Detection. Tao et al. [44] introduce "*Multimodal Consistency Suppression Factor for Fake News Detection*," a model that aims to improve reliability and robustness in fake news identification by mitigating inconsistent visual-textual cues. The framework formulates a **consistency suppression factor (CSF)**, which measures and penalizes semantic incongruence between textual and visual features. This mechanism is embedded into a joint transformer encoder that aligns and integrates multimodal signals, while suppressing noise from deceptive correlations. Furthermore, the model incorporates modality-aware dropout and cross-modal contrastive loss to enhance generalization. Extensive experiments on two widely used multimodal misinformation datasets (Weibo and Twitter) confirm the method's robustness and accuracy under adversarial attacks and real-world distribution shifts. The work addresses the challenges of reliability, adversarial robustness, and explainable inconsistency modeling in safety-critical multimodal applications.

3 Conclusion and Future Outlook

Explainable, robust, and scalable multimodal learning is becoming increasingly central to the development of intelligent systems that must process complex, heterogeneous data across domains. This Special Issue brings together 21 papers that address key challenges in multimodal representation, generation, retrieval, and explainability, showcasing both methodological innovations and practical applications.

In semantic understanding, the selected works highlight progress in contrastive learning, structure-aware modeling, and in-context tuning for robust cross-modal alignment. These advances improve generalization in tasks such as protein captioning, emotion inference, and sarcasm detection under limited supervision.

For vision synthesis, contributions expand the capabilities of diffusion models, GANs, and editing pipelines through prompt engineering, 3D priors, and style disentanglement. These methods demonstrate scalable and controllable generation across domains like fashion design and avatar creation, while also addressing prompt reliability and compositional control.

In multimodal retrieval and recognition, new attention mechanisms and domain-adaptive architectures enhance robustness in video-text alignment, vehicle and person ReID, and hyperspectral classification. These models show strong performance under unsupervised or weakly labeled settings, emphasizing structural fidelity and fine-grained reasoning.

In terms of explainability and trust, the issue explores consistency-aware modeling, norm-grounded reasoning, and frame-based knowledge for socially aligned applications such as fake news detection and dialogue systems. These works highlight the importance of transparency, user alignment, and responsible deployment.

Looking forward, advancing multimodal systems will require scalable, noise-resilient, and interpretable models that bridge foundation capabilities with real-world demands. As vision-language models, generative AI, and multimodal agents evolve, future research will need to focus on instruction tuning, structural representation, and causal reasoning to enable human-aligned, mission-critical applications.

Acknowledgments

We would like to thank all authors who submitted manuscripts to this special issue and all reviewers for devoting a substantial amount of their time to provide high-quality assessments of the submissions' merits. In addition, we would like to express our sincere gratitude to the journal's Editors-in-Chief Prof. Abdulmotaleb El Saddik and Editorial Assistant Ms. Arriane Bustillo for their invaluable support and dedicated service. Special thanks go to Dr. Hao Fei and Asso Prof. Wei Ji for initiating this special issue, and to Prof. Yinwei Wei and Prof. Zhedong Zheng for their enthusiastic responses and successful promotions of the special issue planning and execution. We also greatly appreciate the selfless cooperation from Prof. Jerry Jialie Shen, Prof. Alan Hanjalic, and Prof. Roger Zimmermann, as well as the diligent efforts from all guest editors and reviewers.

Hao Fei and Wei Ji
National University of Singapore, Singapore, Singapore

Yinwei Wei
Monash University, Melbourne, Australia

Zhedong Zheng
University of Macau, Taipa, China

Jialie Shen
City St George's, University of London, London, UK

Alan Hanjalic
Delft University of Technology, Delft, Netherlands

Roger Zimmermann
National University of Singapore, Singapore, Singapore

References

[1] Haoyu Cai, Wenqi Lou, Chao Wang, and Xuehai Zhou. 2024. Picasso: Analyzing prompt design for text-to-image generative diffusion models from a temporal-spatial perspective. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–24. DOI: https://doi.org/10.1145/3724122

[2] Feiyu Chen, Junjie Wang, Yinwei Wei, Hai-Tao Zheng, and Jie Shao. 2022. Breaking isolation: Multimodal graph fusion for multimedia recommendation by edge-wise modulation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 385–394.

[3] Tao Chen, Enwei Zhang, Yuting Gao, Ke Li, Xing Sun, Yan Zhang, Hui Li, and Rongrong Ji. 2024. MMICT: Boosting multi-modal fine-tuning with in-context examples. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–17. DOI: https://doi.org/10.1145/3688804

[4] Hui Cui, Lei Zhu, Jingjing Li, Yang Yang, and Liqiang Nie. 2019. Scalable deep hashing for large-scale social image retrieval. *IEEE Transactions on Image Processing* 29 (2019), 1271–1284.

[5] Kai Cui, Shenghao Liu, Wei Feng, Xianjun Deng, Liangbin Gao, Minmin Cheng, Hongwei Lu, and Laurence T. Yang. 2024. Correlation-aware cross-modal attention network for fashion compatibility modeling in UGC systems. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–24. DOI: https://doi.org/10.1145/3698772

[6] Licun Dai, Zhiming Luo, Yongguo Ling, Jiaxing Chai, and Shao-Zi Li. 2024. Dual-modality-shared learning and label refinement for unsupervised visible-infrared person ReID. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–24. DOI: https://doi.org/10.1145/3724397

[7] Yali Du, Yinwei Wei, Wei Ji, Fan Liu, Xin Luo, and Liqiang Nie. 2023. Multi-queue momentum contrast for microvideo-product retrieval. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, 1003–1011.

[8] Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 5980–5994.

[9] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*.

[10] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022. LasUIE: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

[11] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. VITRON: A unified pixel-level vision LLM for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems*.

[12] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[13] Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. 2025. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the International Conference on Machine Learning*.

[14] Yuan Gan, Ruijie Quan, and Yawei Luo. 2024. ExpAvatar: High-fidelity avatar generation of unseen expressions with 3D face priors. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–21. DOI: https://doi.org/10.1145/3700770

[15] Bingwen Hu, Ping Liu, Zhedong Zheng, and Mingwu Ren. 2022. SPG-VTON: Semantic prediction guidance for multi-pose virtual try-on. *IEEE Transactions on Multimedia* 24 (2022), 1233–1246.

[16] Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. 2020. Unsupervised eyeglasses removal in the wild. *IEEE Transactions on Cybernetics* 51, 9 (2020), 4373–4385.

[17] Zhikun Huang, Zhedong Zheng, Chenggang Yan, Hongtao Xie, Yaoqi Sun, Jianzhong Wang, and Jiyong Zhang. 2021. Real-world automatic makeup via identity preservation makeup net. In *Proceedings of the International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence.

[18] Wei Ji, Long Chen, Yinwei Wei, Yiming Wu, and Tat-Seng Chua. 2022. MRTNet: Multi-resolution temporal network for video sentence grounding. arXiv:2212.13163. Retrieved from https://arxiv.org/abs/2212.13163

[19] Wei Ji, Xi Li, Fei Wu, Zhijie Pan, and Yueting Zhuang. 2019. Human-centric clothing segmentation via deformable semantic locality-preserving network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4837–4848.

[20] Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. 2021. VidVRD 2021: The third grand challenge on video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4779–4783.

[21] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. 2023. Are binary annotations sufficient? Video moment retrieval via hierarchical uncertainty-based active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[22] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3487–3495.

[23] Leyang Jin, Wei Ji, Tat-seng Chua, and Zhedong Zheng. 2025. Coarse-to-fine cross-modality generation for enhancing vehicle re-identification with high-fidelity synthetic data. In *Proceedings of the International Conference on Robotics and Automation*.

[24] Weipeng Jing, Peilun Kang, Donglin Di, Juntao Gu, Linhui Li, Mahmoud Emam, Linda Mohaisen, Xun Yang, and Chao Li. 2024. SRF: SpectrumRecombineFormer for hyperspectral image classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–25. DOI: https://doi.org/10.1145/3715698

[25] Bobo Li, Hao Fei, Fei Li, Tat-Seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–19. DOI: https://doi.org/10.1145/3689646

[26] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5923–5934.

[27] Haoran Li, Yanbin Hao, Jiarui Yu, Bin Zhu, Shuo Wang, and Tong Xu. 2024. CVLP-NaVD: Contrastive visual-language pre-training models for non-annotated visual description. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–23. DOI: https://doi.org/10.1145/3708348

[28] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4654–4662.

[29] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

[30] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing* 31 (2022), 3780–3792.

[31] Jinliang Liu, Zhedong Zheng, Zongxin Yang, and Yi Yang. 2024. High fidelity makeup via 2D and 3D identity preservation net. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 8 (2024), 1–24.

[32] Yu Liu, Haipeng Chen, Guihe Qin, Jincai Song, and Xun Yang. 2024. Bias mitigation and representation optimization for noise-robust cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–17. DOI: https://doi.org/10.1145/3700596

[33] Yaxin Liu, Jianlong Wu, Leigang Qu, Tian Gan, Jianhua Yin, and Liqiang Nie. 2022. Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia* (2022).

[34] Yiwei Ma, Yijun Fan, Jiayi Ji, Haowei Wang, Haibing Yin, Xiaoshuai Sun, and Rongrong Ji. 2024. Creating high-quality 3D content by bridging the gap between text-to-2D and text-to-3D generation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–23. DOI: https://doi.org/10.1145/3687475

[35] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

[36] Shilin Qu, Weiqing Wang, Xin Zhou, Haolan Zhan, Zhuang Li, Lizhen Qu, Linhao Luo, Yuan-Fang Li, and Gholamreza Haffari. 2024. Scalable frame-based construction of sociocultural norm bases for socially aware dialogues. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–17. DOI: https://doi.org/10.1145/3697838

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

[38] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3654–3663.

[39] Fei Shen, Xiaoyu Du, Liyan Zhang, Xiangbo Shu, and Jinhui Tang. 2024. Triplet contrastive representation learning for unsupervised vehicle re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–23. DOI: https://doi.org/10.1145/3695255

[40] Leqi Shen, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. 2024. Spatio-temporal attention for text-video retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–20. DOI: https://doi.org/10.1145/3715137

[41] Zhengwentai Sun, Yanghong Zhou, and Tracy Mok. 2024. CoDE-GAN: Content decoupled and enhanced GAN for sketch-guided flexible fashion editing. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–24. DOI: https://doi.org/10.1145/3712063

[42] Yucheng Suo, Zhedong Zheng, Xiaohan Wang, Bang Zhang, and Yi Yang. 2024. Jointly harnessing prior structures and temporal consistency for sign language video generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 6 (2024), 1–18.

[43] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5100–5111.

[44] Zhulin Tao, Runze Zhao, Xin Shi, Xingyu Gao, Xi Wang, and Xianglin Huang. 2024. Multimodal consistency suppression factor for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–19. DOI: https://doi.org/10.1145/3699959

[45] Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. 2024. S^3 agent: Unlocking the power of VLLM for zero-shot multi-modal sarcasm detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–16. DOI: https://doi.org/10.1145/3690642

[46] Xiaodong Wang, Zhedong Zheng, Yang He, Fei Yan, Zhiqiang Zeng, and Yi Yang. 2023. Progressive local filter pruning for image retrieval acceleration. *IEEE Transactions on Multimedia* 25 (2023), 9597–9607.

[47] Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. 2022. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia* 25 (2022), 6079–6089.

[48] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv:2503.12605. Retrieved from https://arxiv.org/abs/2503.12605

[49] Hongchen Wei and Zhenzhong Chen. 2024. Improving domain generalization for image captioning with unsupervised prompt learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–23. DOI: https://doi.org/10.1145/3715136

[50] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 2651–2658.

[51] Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. Information screening whilst exploiting! Multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14734–14751.

[52] Shengqiong Wu, Hao Fei, and Tat-Seng Chua. 2025. Universal scene graph generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14158–14168.

[53] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Towards semantic equivalence of tokenization in multimodal LLM. arXiv:2406.05127. Retrieved from https://arxiv.org/abs/2406.05127

[54] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, 53366–53397.

[55] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. 2023. Imagine that! Abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 79240–79259.

[56] Ying Wu, Qihe Pan, Zhen Zhao, Zicheng Wang, Sifan Long, and Ronghua Liang. 2024. SOEDiff: Efficient distillation for small object editing. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–19. DOI: https://doi.org/10.1145/3715915

[57] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[58] Shuyu Yang, Yaxiong Wang, Yongrui Li, Li Zhu, and Zhedong Zheng. 2025. Minimizing the pretraining gap: Domain-aligned text-based person retrieval. arXiv:2507.10195. Retrieved from https://arxiv.org/abs/2507.10195

[59] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–10.

[60] Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022. Structured and natural responses co-generation for conversational search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 155–164.

[61] Hang Yu, Jiahao Wen, and Zhedong Zheng. 2025. CAMeL: Cross-modality adaptive meta-learning for text-based person retrieval. *IEEE Transactions on Information Forensics and Security* (2025).

[62] Xuzheng Yu, Tian Gan, Yinwei Wei, Zhiyong Cheng, and Liqiang Nie. 2020. Personalized item recommendation for second-hand trading platform. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3478–3486.

[63] Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. 2024. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. arXiv:2410.11236. Retrieved from https://arxiv.org/abs/2410.11236

[64] Jianrong Zhang, Hehe Fan, and Yi Yang. 2024. Protein captioning: Bridging the gap between protein sequences and natural languages. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–23. DOI: https://doi.org/10.1145/3705322

[65] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Yi Yang, and Tat-Seng Chua. 2023. Multi-view consistent generative adversarial networks for compositional 3D-aware image synthesis. *International Journal of Computer Vision* 131, 8 (2023), 2219–2242.

[66] Yue Zhang, Chao Wang, Fei Fang, Yunzhi Zhuge, Hehe Fan, Xiaojun Chang, Cheng Deng, and Yi Yang. 2024. SAMControl: Controlling pose and object for image editing with soft attention mask. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–28. DOI: https://doi.org/10.1145/3702999

[67] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. 2023. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 29th ACM International Conference on Multimedia*.

[68] Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 7960–7977.

[69] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. 2020. VehicleNet: Learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia* 23 (2020), 2683–2693.

[70] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. 2022. Parameter-efficient person re-identification in the 3D space. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[71] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2138–2147.

[72] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 2 (2020), 1–23.

[73] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 3754–3762.

[74] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

[75] Xiaojie Zhou, Hang Yu, Shengjie Yang, Jing Huo, and Pinzhuo Tian. 2024. Learning from orthogonal space with multimodal large models for generalized few-shot segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 11 (2025), 1–22. DOI: https://doi.org/10.1145/3712597

[76] Yinan Zhou, Yaxiong Wang, Haokun Lin, Chen Ma, Li Zhu, and Zhedong Zheng. 2025. Scale up composed image retrieval learning via modification text generation. *IEEE Transactions on Multimedia* (2025).