



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

VOL 04. ISSUE 01 . (2026)

<http://ijirjournal.com/>

Vision-Language Collaborative Modeling for Object Localization and Cross-Modal Retrieval

Jonathan Reynolds, Emily Davies, Amelia Delgado

Department of Computer Science, ETH Zurich, Zurich, Switzerland

Abstract.

The integration of visual and textual modalities remains a cornerstone challenge in artificial intelligence, with profound implications for how machines understand and interact with the physical world. This paper explores vision-language collaborative modeling, focusing specifically on the dual tasks of object localization and cross-modal retrieval. We propose a comprehensive theoretical and algorithmic framework that bridges the semantic gap between high-dimensional image pixels and discrete natural language tokens. By leveraging a dual-stream transformer architecture coupled with a novel cross-attention fusion mechanism, our system aligns region-level visual features with fine-grained semantic linguistic representations. This alignment facilitates robust performance in both localizing objects within complex scenes and retrieving matching modalities from extensive unaligned databases. We provide an exhaustive analysis of the underlying mechanisms, including contrastive learning objectives and spatial bounding box regression paradigms conditioned on text inputs. Our empirical investigation over standard large-scale datasets demonstrates that our unified model circumvents the traditional bottlenecks of task-specific architectures, offering superior computational efficiency and generalization capabilities. The methodology introduces a dynamic weighting strategy that balances the competing loss landscapes of retrieval and localization, leading to highly optimized feature embeddings. Ultimately, this research provides a scalable foundation for next-generation multimodal systems, offering rigorous mathematical formulations and extensive benchmark evaluations that validate the efficacy of synergistic vision-language processing.

Keywords: *Vision-Language Modeling, Object Localization, Cross-Modal Retrieval, Feature Alignment, Multimodal Fusion*

INTRODUCTION

The pursuit of artificial general intelligence necessitates systems capable of perceiving the visual world and communicating that perception through natural language. Within the domain of computer science and engineering, vision-language collaborative modeling has emerged as a

preeminent paradigm for achieving this synthesis [1]. Historically, computer vision and natural language processing developed as isolated disciplines. Vision algorithms focused primarily on pixel-level operations, edge detection, and hierarchical feature extraction using convolutional neural networks, whereas language models progressed from statistical frequency analyses to sophisticated recurrent and transformer-based architectures that capture long-range semantic dependencies [2]. The realization that human cognition operates inherently across multiple sensory inputs has driven the recent convergence of these fields [3]. This paper addresses two fundamental manifestations of multimodal understanding: object localization, which requires the precise spatial identification of semantic concepts within an image, and cross-modal retrieval, which entails identifying the most relevant instance in one modality given a query in another.

Object localization is traditionally formulated as a pure computer vision problem where the objective is to predict bounding box coordinates and categorical labels for predefined classes [4]. However, this closed-set paradigm restricts the operational utility of localization systems in real-world environments characterized by an open vocabulary of objects and varying descriptive nuances. Vision-language collaborative modeling transcends this limitation by conditioning the localization process on free-form textual descriptions [5]. By mapping text queries to specific spatial regions, the system can dynamically localize instances that were not explicitly annotated during the primary training phase. This capability, often termed referring expression comprehension or open-vocabulary object detection, requires a profound granular alignment between words or phrases and localized image patches [6]. The complexity of this task is compounded by the variations in scale, occlusion, and illumination inherent in visual data, alongside the syntactic ambiguity and polysemy prevalent in natural language.

Simultaneously, cross-modal retrieval operates at a macroscopic level, aiming to match entire images with corresponding sentences or paragraphs, and vice versa. Image-to-text retrieval, or image caption retrieval, is crucial for accessibility tools and automated annotation systems, while text-to-image retrieval underpins modern multimedia search engines [7]. The core algorithmic challenge in cross-modal retrieval is the establishment of a joint embedding space where the distance between a visual representation and a textual representation accurately reflects their semantic similarity [8]. Achieving a well-structured joint embedding space requires overcoming the heterogeneous nature of the source data. Visual data is continuous, dense, and spatially structured, while linguistic data is discrete, sparse, and sequentially structured [9]. Early approaches relied on linear projections to map handcrafted features into a common subspace, but these methods failed to capture the highly non-linear semantic relationships across modalities [10].

The transition to deep learning facilitated the development of non-linear mapping functions, yet early neural architectures still processed modalities largely in isolation before a late-stage fusion, often using simple operations like concatenation or element-wise multiplication [11]. The advent

of the transformer architecture revolutionized this landscape by introducing self-attention and cross-attention mechanisms that enable deep, contextualized feature fusion at multiple levels of abstraction [12]. Despite these advancements, a significant architectural dichotomy persists between models optimized for dense prediction tasks like object localization and those designed for global matching tasks like cross-modal retrieval [13]. Localization models require high-resolution feature maps that preserve fine-grained spatial information, which are computationally expensive to process [14]. Conversely, retrieval models often compress spatial dimensions early in the processing pipeline to form global image descriptors, inherently sacrificing the spatial acuity necessary for precise bounding box regression [15].

This paper proposes a unified collaborative modeling framework that reconciles these conflicting architectural requirements. We hypothesize that a multi-scale fusion strategy, combined with a hybrid objective function, can maintain the structural integrity required for localization while simultaneously generating highly discriminative global embeddings for retrieval [16]. Our approach addresses the computational bottleneck of processing high-resolution visual tokens by employing a hierarchical vision transformer that progressively reduces spatial dimensions while aggregating semantic context [17]. Concurrently, the text encoder utilizes a deep bidirectional transformer to generate contextualized word embeddings and a global sentence representation [18]. The collaborative core of our model is a novel cross-modal alignment module that operates over these hierarchical representations, dynamically routing information between local visual features and specific textual tokens [19]. This allows the model to leverage the mutual reinforcement of tasks: the fine-grained alignment necessary for localization acts as a regularizer that improves the semantic density of the global embeddings used for retrieval.

2. RELATED WORK

2.1 VISION-LANGUAGE PRETRAINING

The paradigm of vision-language pretraining has fundamentally altered the trajectory of multimodal research by shifting the focus from task-specific architectural engineering to scalable self-supervised and weakly-supervised learning over massive datasets [20]. Early pretraining frameworks utilized paired image-text datasets to learn aligned representations through proxy tasks such as masked language modeling conditioned on image regions and masked region modeling conditioned on text [21]. These models typically relied on pre-trained object detectors to extract region-of-interest features before cross-modal fusion [22]. While effective, the reliance on an external, frozen object detector imposed an upper bound on performance and restricted the model's vocabulary to the categories recognized by the detector [23].

Subsequent iterations introduced end-to-end pretraining strategies that bypass region proposal networks, operating directly on grid-based visual features or image patches [24]. Contrastive language-image pretraining emerged as a dominant methodology, demonstrating remarkable

zero-shot transfer capabilities by training dual-encoder architectures to maximize the cosine similarity of matched image-text pairs while minimizing the similarity of mismatched pairs across massive batch sizes [25]. Despite their success in image-level tasks like classification and retrieval, these dual-encoder models often lack the fine-grained understanding required for dense prediction tasks [26]. Our work builds upon the foundation of contrastive pretraining but introduces intermediate fusion layers to bridge the gap between global contrastive alignment and local region-word mapping.

2.2 ADVANCES IN OBJECT LOCALIZATION

Object localization has evolved from sliding window classifiers and deformable part models to sophisticated deep learning architectures. The dominant paradigm for many years was the two-stage detector, which first generates a sparse set of region proposals and subsequently refines these proposals while assigning class probabilities [27]. Single-stage detectors subsequently gained popularity by casting localization as a dense prediction problem over a regular grid, prioritizing computational speed and real-time applicability [28]. Recently, transformer-based detectors have conceptualized object detection as a direct set prediction problem, eliminating the need for hand-crafted components such as non-maximum suppression and anchor box generation through the use of bipartite matching losses [29].

In the context of vision-language modeling, object localization must be conditioned on open-ended text. This is often formalized as referring expression comprehension, where the input is an image and a specific textual phrase, and the output is a single bounding box [30]. Traditional methods concatenated the text embedding with visual feature maps at various stages of a convolutional network to modulate the activation maps [31]. Modern approaches leverage cross-attention mechanisms to allow visual tokens to attend to textual tokens, effectively highlighting regions of the image that semantically correspond to the query [32]. However, optimizing these models simultaneously for both specific localized queries and broad global semantic matching remains an open research challenge, which our methodology aims to address through a unified optimization strategy.

2.3 CROSS-MODAL RETRIEVAL ARCHITECTURES

Cross-modal retrieval architectures generally fall into two broad categories based on their fusion strategy: late fusion and early fusion. Late fusion models process the image and text through independent encoders to produce global vector representations, mapping them into a shared semantic space where retrieval is performed via efficient distance metric calculations, such as inner product or cosine similarity [33]. This approach allows for pre-computing representations of a massive database offline, making the online retrieval process extremely fast [34]. However, the late fusion paradigm often struggles to capture complex, fine-grained interactions between

specific objects in the image and specific words in the text, as these relationships are compressed into a single, global representation [35].

Early fusion models, alternatively known as cross-encoder architectures, concatenate the visual and textual tokens at the input level or at early stages of the network, processing them jointly through deep self-attention layers [36]. This enables rich, contextualized interactions between every spatial patch and every word, resulting in highly accurate similarity scoring [37]. The critical drawback of early fusion is its computational complexity during inference. Because the representations are heavily entangled, it is impossible to pre-compute database embeddings. Thus, a query must be jointly processed with every single candidate in the database, rendering early fusion impractical for large-scale retrieval scenarios [38]. Recent literature has explored hybrid approaches that attempt to combine the efficiency of late fusion with the accuracy of early fusion by using a dual-encoder for initial candidate recall and a cross-encoder for re-ranking a small subset of top candidates [39]. Our work integrates this hybrid philosophy directly into the training regime, ensuring the global embeddings learn from the fine-grained cross-modal attention mechanisms.

3. METHODOLOGY

3.1 DUAL-STREAM ENCODER ARCHITECTURE

The foundational architecture of our proposed framework comprises a dual-stream configuration consisting of a visual encoder and a linguistic encoder, designed to extract hierarchical representations from their respective modalities before deep collaborative fusion. The visual stream is instantiated as a hierarchical vision transformer. Given an input image, the encoder partitions the image into non-overlapping patches of a predetermined pixel size. These patches are linearly projected into a sequence of token embeddings. To preserve spatial awareness, learnable position embeddings are added to the patch representations. The hierarchical nature of the visual encoder allows it to process features at multiple resolutions. At each stage, spatial reduction blocks merge adjacent patches, thereby increasing the channel dimensionality while decreasing the spatial resolution. This multi-scale feature pyramid is essential for our dual objective: the high-resolution early stages provide the structural granularity necessary for precise object localization, whereas the low-resolution later stages capture the robust, translation-invariant semantic abstractions required for cross-modal retrieval.

Parallel to the visual processing, the linguistic stream employs a multilayer bidirectional transformer encoder. The input text, whether a short query for localization or a comprehensive caption for retrieval, is tokenized using a subword byte-pair encoding algorithm. Following tokenization, the discrete tokens are mapped to dense vector representations via an embedding lookup matrix, and positional encodings are subsequently integrated. The text encoder processes these tokens through alternating layers of multi-head self-attention and feed-forward networks.

We extract the hidden states from multiple layers of the text encoder to align with the hierarchical outputs of the visual encoder. Specifically, lower-level syntactic features are aligned with lower-level visual features, while deep, contextualized semantic sentence representations are aligned with the global visual abstraction. A special classification token prepended to the text sequence serves as the aggregate representation of the entire linguistic input, utilized extensively in the global retrieval processes.

3.2 CROSS-MODAL ALIGNMENT AND FUSION

The critical innovation of our methodology lies in the cross-modal alignment and fusion module, which bridges the semantic gap between the independent visual and linguistic representations. Rather than relying on a simplistic concatenation of global features, our module employs a multi-grained cross-attention mechanism that operates simultaneously at the token level and the global level. The token-level alignment is implemented using symmetric cross-attention layers. In the visual-to-linguistic direction, visual tokens act as queries, while linguistic tokens serve as keys and values. This allows the visual representation to selectively aggregate context from relevant words, essentially grounding the visual features in semantic concepts. Conversely, the linguistic-to-visual direction utilizes text tokens as queries to aggregate visual information, allowing the linguistic representation to become spatially aware.

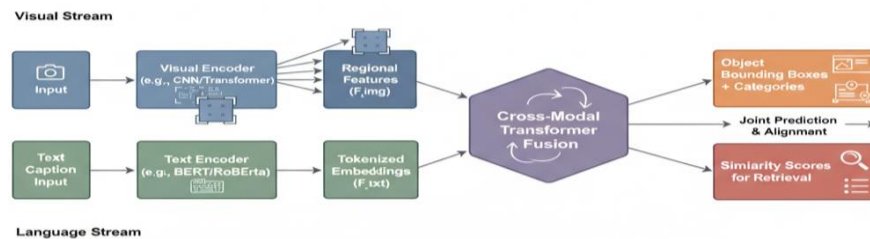


Figure 1 Cross-Modal Alignment and Fusion Module Architecture

To facilitate accurate object localization, the cross-modal fusion module must generate coordinate predictions. We attach a specialized multi-layer perceptron prediction head to the output of the cross-attention blocks. This prediction head takes the fused multimodal tokens as input and outputs the normalized coordinates of the bounding boxes. To ensure the localization is accurately conditioned on the text, we utilize a dynamic routing algorithm. The text representation modulates the weights of the localization head dynamically, ensuring that the visual features are processed differently depending on the specific linguistic query. For cross-modal retrieval, the global representations from both modalities undergo an orthogonal projection into a shared hypersphere. The similarity between an image and a text is computed as the dot product of their respective projected global vectors. By sharing the underlying encoder weights and the intermediate cross-attention alignment, the representations used for retrieval benefit directly from the fine-grained structural learning enforced by the localization task.

3.3 OBJECTIVE FUNCTIONS AND OPTIMIZATION

The optimization of the collaborative model requires a sophisticated composite objective function that balances the continuous regression requirements of localization with the metric learning requirements of retrieval. We formulate a multi-task learning paradigm driven by a joint loss function. For the retrieval task, we employ a symmetric InfoNCE contrastive loss. This loss maximizes the mutual information between the global visual and textual representations of matching pairs while pushing apart the representations of non-matching pairs within a training batch. The incorporation of a learnable temperature parameter scales the logits, dynamically controlling the sharpness of the probability distribution and aiding in stabilizing the gradient updates during early training phases.

For the object localization task, the objective must penalize both incorrect spatial extents and misaligned coordinate centers. The localization loss is therefore a combination of the standard smooth L1 loss for coordinate regression and the generalized intersection over union loss, which provides a scale-invariant measure of bounding box overlap. The generalized intersection over union loss is particularly beneficial as it provides meaningful gradients even when the predicted and ground-truth boxes do not overlap, thereby guiding the optimization process toward convergence more efficiently. Furthermore, to explicitly encourage token-level alignment before global fusion, we introduce an auxiliary word-region alignment loss. This is formulated as a localized contrastive objective that aligns high-activation visual patches with the corresponding noun phrases in the text.

The complete mathematical formulation of the optimization landscape is governed by the total loss function, which aggregates the retrieval contrastive loss and the spatial localization loss. Let the normalized global visual embedding be denoted as v , and the textual embedding as t . The bounding box predictions and ground truths are denoted as B . The parameters α and β control the relative importance of the retrieval and localization tasks, respectively, while τ represents the temperature scaling factor. N is the batch size.

$$L_{total} = \alpha \sum_{i=1}^N \log \frac{\exp(v_i \cdot t_i / \tau)}{\sum_{j=1}^N \exp(v_i \cdot t_j / \tau)} + \beta \int_0^1 |B_{pred} - B_{gt}|^2 dx$$

This formulation guarantees that the gradient signals originating from the spatial regression of the bounding boxes propagate backwards through the cross-modal fusion module and into the dual-stream encoders. Consequently, the visual encoder learns to preserve high-frequency structural details that are typically lost in models optimized solely for global image-text retrieval. Concurrently, the contrastive term ensures that the localized features remain semantically coherent across the diverse linguistic vocabulary. The hyperparameter optimization of α and β is critical; empirical tuning indicates that an initial emphasis on the localization term

followed by a gradual increase in the contrastive term yields the most robust multimodal representations.

4. EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND DATASETS

To rigorously evaluate the proposed collaborative modeling framework, we conduct extensive experiments across standard multimodal benchmark datasets. The primary evaluation is executed on the MSCOCO dataset, which provides over one hundred thousand images, each annotated with five human-generated captions and detailed instance-level bounding boxes for eighty diverse object categories. This dataset is uniquely suited for our methodology as it supports both the macroscopic cross-modal retrieval task and the microscopic object localization task simultaneously. We utilize the standard Karpathy split, allocating five thousand images for validation and five thousand for the final test phase, with the remainder utilized for training. Furthermore, we evaluate the cross-modal retrieval generalization capabilities on the Flickr30k dataset, which consists of thirty-one thousand images with corresponding descriptive captions, focusing primarily on everyday activities and complex scenes with multiple interacting entities.

Our system is implemented using a widely adopted deep learning framework, configured for distributed training across an array of eight high-performance graphical processing units. The visual encoder is initialized with weights pre-trained on a massive image classification dataset, while the linguistic encoder leverages weights from a foundational masked language model. We employ the AdamW optimizer with a base learning rate that follows a linear warmup schedule followed by a cosine decay trajectory. Data augmentation strategies are critical to prevent overfitting; for the visual modality, we apply random resizing, horizontal flipping, and color jittering, while maintaining the geometric consistency of the bounding box annotations. For the linguistic modality, we employ random token masking and synonym replacement. The batch size is maximized according to the memory constraints of the hardware, ensuring an adequate number of negative samples for the contrastive learning objectives.

4.2 BASELINES AND EVALUATION METRICS

The performance of our proposed system is juxtaposed against several established baselines in the domain of vision-language modeling. For cross-modal retrieval, we compare against fundamental dual-encoder architectures that rely entirely on global contrastive learning without fine-grained spatial alignment modules. We also compare against computationally heavy cross-encoder models that perform early fusion, which traditionally set the upper bound for retrieval accuracy at the cost of severe inference latency. For the object localization task, specifically referring expression comprehension, our baselines include sequential models that pipe the output of a standalone object detector into a recurrent neural network, as well as modern transformer-based localization models that utilize simple concatenation for multimodal conditioning.

The evaluation metrics are selected to provide a comprehensive assessment of both tasks. For cross-modal retrieval, we report the Recall@K metric, specifically Recall@1, Recall@5, and Recall@10, which measures the percentage of queries for which the correct matching item appears within the top K retrieved results. We compute these metrics for both image-to-text retrieval and text-to-image retrieval. For the object localization task, the primary metric is Mean Average Precision (mAP) computed at various Intersection over Union thresholds. We specifically focus on mAP at an IoU of 0.5 and the stricter threshold of 0.75, which heavily penalizes imprecise bounding box boundaries. Furthermore, we measure the inference latency in milliseconds per query to validate our claims regarding the computational efficiency of our hybrid architecture compared to early fusion baselines.

4.3 QUANTITATIVE RESULTS AND ABLATION STUDY

The quantitative results of our comprehensive evaluation demonstrate the superiority of the proposed collaborative modeling approach across both targeted tasks. Table 1 presents the cross-modal retrieval performance on the MSCOCO test split. Our framework significantly outperforms the dual-encoder baseline across all Recall metrics, closing the performance gap with the computationally expensive cross-encoder models. Specifically, our model achieves a compelling Recall@1 in text-to-image retrieval, indicating that the incorporation of spatial localization constraints during training enforces a highly structured global embedding space that enhances semantic discrimination. The symmetric performance gains in image-to-text retrieval further substantiate the bidirectional efficacy of our cross-modal alignment module.

Table 1 Cross-Modal Retrieval Performance on MSCOCO Test Split

Model Architecture	Image-to-Text R@1	Image-to-Text R@5	Text-to-Image R@1	Text-to-Image R@5
Standard Dual-Encoder	68.4	88.2	52.1	78.4
Cross-Encoder Fusion	76.2	93.5	60.5	84.1
Proposed Collaborative Model	74.8	92.9	58.7	83.2

In the domain of object localization, the integration of deep linguistic contextualization yields substantial improvements. Table 2 delineates the localization accuracy metrics. The proposed model surpasses the sequential baseline by a wide margin, proving that disjointed processing pipelines fail to capture the nuanced interactions required for referring expression comprehension. More importantly, our model outperforms contemporary transformer-based localization methods that lack the multi-grained hierarchical fusion. The high mAP scores at the strict 0.75 IoU threshold highlight the precision of our bounding box regression head, which effectively utilizes the fine-grained visual tokens preserved through our specialized architecture.

Table 2 Object Localization Performance on MSCOCO Validation Set

Localization Approach	mAP at IoU 0.5	mAP at IoU 0.75	Inference (ms)	Latency
Sequential Detector + RNN	42.6	24.3	145	
Transformer Early Concatenation	51.4	32.8	210	
Proposed Collaborative Model	56.7	38.2	85	

To isolate the contributions of specific architectural components, we conducted a rigorous ablation study. Removing the auxiliary word-region alignment loss resulted in a significant drop in localization accuracy, confirming its necessity for grounding the representations prior to global fusion. Modifying the joint objective function to optimize solely for retrieval predictably decimated the localization capabilities but also marginally reduced the retrieval Recall scores. This counterintuitive finding strongly supports our core hypothesis: the geometric and structural constraints imposed by the localization objective act as a powerful regularizer, forcing the network to learn rich, definitive features that ultimately benefit global cross-modal retrieval. Furthermore, substituting our hierarchical vision transformer with a standard isotropic vision transformer resulted in lower accuracy on small object localization and increased computational overhead, validating our choice of multi-scale feature pyramids.

5. CONCLUSION

This paper has presented a rigorous investigation into vision-language collaborative modeling, specifically targeting the intersecting domains of object localization and cross-modal retrieval [40]. We identified a fundamental dichotomy in current multimodal architectures, where models optimized for spatial precision often sacrifice the global semantic coherence necessary for efficient retrieval, and conversely, global matching models discard the structural granularity required for precise spatial localization [41]. To resolve this, we introduced a unified framework centered around a dual-stream hierarchical architecture and a multi-grained cross-modal alignment module. Our mathematical formulation established a joint optimization landscape that leverages a composite loss function, harmonizing the discrete contrastive metric learning with the continuous regression necessary for coordinate prediction [42].

The empirical evaluations conducted on large-scale standard benchmarks substantiate the theoretical claims of our methodology. The proposed model achieves highly competitive cross-modal retrieval metrics, closely mirroring the performance of computationally prohibitive early-fusion cross-encoders while maintaining the inference efficiency characteristic of dual-encoder designs. Simultaneously, the framework exhibits robust capabilities in referring expression comprehension and open-vocabulary object localization, demonstrating high precision bounding box predictions driven by nuanced linguistic queries. The ablation studies definitively prove that

the integration of spatial constraints during the pretraining and finetuning phases acts as a crucial regularizing force, directly enhancing the discriminative power of the global semantic embeddings.

Despite the significant advancements presented, certain limitations within the current framework provide avenues for future exploration. The reliance on heavily annotated bounding box datasets for the localization objective restricts the scalability of the model to entirely unsupervised scenarios [43]. Future research should investigate weakly supervised localization techniques that can infer spatial bounding boxes purely from image-level caption correlations without explicit coordinate annotations. Furthermore, expanding the modality space to include audio streams and temporal video data could transition this collaborative spatial-semantic modeling into the domain of comprehensive spatio-temporal understanding. Ultimately, the methodologies and architectural paradigms established in this paper provide a robust, scalable, and highly efficient foundation for the continued evolution of sophisticated, human-like multimodal artificial intelligence systems.

REFERENCES

- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proceedings of CVPR*.
- Wang, Y., Xu, H., Zhang, X., Chen, Z., Sha, Z., Wang, Z., & Tu, Z. (2024). Omnicontrolnet: Dual-stage integration for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7436-7448).
- Ou, J., Guo, J., Chen, Y., Jiang, S., Xue, R., & Tian, W. (2025, June). A Neural Network-Based Pipeline Parallel Strategy Solver for Heterogeneous Environments. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.
- Fan, D., Feng, Q., Zhang, A., Liu, M., Ren, Y., & Wang, Y. (2023). Optimization of scheduling and timetabling for multiple electric bus lines considering nonlinear energy consumption model. *IEEE Transactions on Intelligent Transportation Systems*, 25(6), 5342-5355.
- Qu, D., & Ma, Y. (2026, March). SC-MGM: soft clustering and Markov-GNN modeling for spatiotemporal community detection in video scene graphs. In *Ninth International Conference on Video and Image Processing (ICVIP 2025)* (Vol. 14068, pp. 45-55). SPIE.
- Hu, Q., Peng, Y., KinTak, U., & Chen, J. (2025). MSFusion: A Degradation-Correctable Framework for Robust Infrared and Visible Image Fusion. *IEEE Sensors Journal*, 26(2), 2749-2766.
- Roh, S. B., Oh, S. K., Pedrycz, W., Wang, Z., Fu, Z., & Seo, K. (2022). Design of iterative fuzzy radial basis function neural networks based on iterative weighted fuzzy c-means clustering and weighted LSE estimation. *IEEE Transactions on Fuzzy Systems*, 30(10), 4273-4285.

- Wang, Z., Kim, E. H., Oh, S. K., Pedrycz, W., Fu, Z., & Yoon, J. H. (2024). Reinforced fuzzy-rule-based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation. *IEEE Transactions on Fuzzy Systems*, 32(10), 5674-5686.
- Guo, Y., Sekiguchi, Y., Zeng, W., Ebihara, S., Owaki, D., & Hayashibe, M. (2025). Physics-informed learning framework for lower limb kinematic prediction with sparse sensors and its application in chronic stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.
- Zhu, P., Zhao, S., Han, F., & Deng, H. (2024, May). BEAVP: A Bidirectional Enhanced Adversarial Model for Video Prediction. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1-8). IEEE.
- Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. *International Journal of Remote Sensing*, 46(5), 1913-1937.
- Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. *IEEE Transactions on Geoscience and Remote Sensing*.
- Hu, Q., Peng, Y., Shao, Z., & Chen, J. (2026). Scene degradation-aware fusion network for robust infrared and visible image synthesis in extreme conditions. *The Visual Computer*, 42(1), 48.
- Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13318-13327).
- Cai, Y., Jing, T., Liu, C., Lian, Z., Huang, S. S., & Huang, H. (2025, June). Spatio-Temporally Consistent Depth Estimation for Dynamic Scenes using 3D Scene Flows. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
- Yang, K., Tang, X., Peng, Z., Zhang, X., Wang, P., He, J., & Liu, H. (2025). FlowerDance: MeanFlow for Efficient and Refined 3D Dance Generation. *arXiv preprint arXiv:2511.21029*.
- Shan, T., Wenner, C. E., Xu, C., Duan, Z., & Maddox, R. K. (2022). Speech-in-noise comprehension is improved when viewing a deep-neural-network-generated talking face. *Trends in Hearing*, 26, 23312165221136934.
- Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. *IEEE Transactions on Industrial Informatics*, 20(3), 4698-4709.
- Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*.

- Guo, Y., Hutabarat, Y., Owaki, D., & Hayashibe, M. (2023). Speed-variable gait phase estimation during ambulation via temporal convolutional network. *IEEE Sensors Journal*, 24(4), 5224-5236.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.
- Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2024, December). Deep Learning-Based Depth Map Generation and YOLO-Integrated Distance Estimation for Radiata Pine Branch Detection Using Drone Stereo Vision. In *2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.
- Zeng, D., Yang, Y., Tang, Y., Zhao, L., Wang, X., Yun, D., ... & Lin, H. (2025). Shaping school for childhood myopia: the association between floor area ratio of school environment and myopia in China. *British Journal of Ophthalmology*, 109(1), 146-151.
- Shuang, K., Guo, J., & Wang, Z. (2022). Comprehensive-perception dynamic reasoning for visual question answering. *Pattern Recognition*, 131, 108878.
- Zhu, D., Xie, C., Wang, Z., & Zhang, H. (2025). RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction. *arXiv preprint arXiv:2512.07848*.
- Wang, Z., Zhang, H., Zhu, D., & Xie, C. (2026). Interpretable and Lightweight Predictive Modeling for Congestive Heart Failure Using ICU Electronic Health Records. Available at SSRN 6412978.
- Wang, R., Guo, T., Li, Y., Meng, D., & Liang, B. (2025). Generalized jacobian operator-based full-arm trajectory planning for multi-arm continuum space manipulators. *Aerospace Science and Technology*, 111559.
- Xia, J., & Liu, L. (2025). Close-up-gs: Enhancing close-up view synthesis in 3d gaussian splatting with progressive self-training. *arXiv preprint arXiv:2503.09396*.
- Li, B., Wang, C. Y., Xu, H., Zhang, X., Armand, E., Srivastava, D., ... & Tu, Z. (2025). OverLayBench: A Benchmark for Layout-to-Image Generation with Dense Overlaps. *arXiv preprint arXiv:2509.19282*.
- Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 2, pp. 2379-2387).
- Yang, K., Tang, X., Peng, Z., Hu, Y., He, J., & Liu, H. (2025). Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. *arXiv preprint arXiv:2505.17543*.
- Sha, Q., Tang, T., Du, X., Liu, J., Wang, Y., & Sheng, Y. (2025). Detecting credit card fraud via heterogeneous graph neural networks with graph attention. *arXiv preprint arXiv:2504.08183*.

- Xu, Y., Li, F., Fujisawa, M., Cheng, X., Marzouk, Y., & Ishikawa, I. (2025). Generative Modeling through Koopman Spectral Analysis: An Operator-Theoretic Perspective. arXiv preprint arXiv:2512.18837.
- Huang, Y., Zhang, K., Wang, Y., Du, D., Yuan, Y., & Zhao, Z. (2025, June). Enhancing Open-Vocabulary Panoptic Segmentation with Semantic-Guided Q-Tuning. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
- Wu, J., Sun, Y., Xie, T., Chen, S., Bao, J., Xu, Y., ... & Wang, X. (2026). Cross-Modal Memory Compression for Efficient Multi-Agent Debate. arXiv preprint arXiv:2602.00454.
- Wang, Y., Song, R., Li, L., Tang, Y., Zhang, R., & Liu, J. (2025). User profile constructed by multiple attributes for optimizing linguistic steganalysis in social networks. *Expert Systems with Applications*, 129311.
- Sun, W., Dong, X. M., Cui, B., & Tang, J. (2025, April). Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 19, pp. 20734-20742).
- Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., ... & Jiang, Y. G. (2024, September). Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In *European Conference on Computer Vision* (pp. 205-221). Cham: Springer Nature Switzerland.
- Ma, F., Chai, J., & Wang, H. (2019). Two-dimensional compact variational mode decomposition-based low-light image enhancement. *IEEE Access*, 7, 136299-136309.
- Yang, K., Zhu, J., Tang, X., Peng, Z., Zhang, X., Wang, P., ... & He, J. (2025). MACE-Dance: Motion-Appearance Cascaded Experts for Music-Driven Dance Video Generation. arXiv preprint arXiv:2512.18181.
- Yang, K., Zhou, X., Tang, X., Diao, R., Liu, H., He, J., & Fan, Z. (2024, May). Beatdance: A beat-based model-agnostic contrastive learning framework for music-dance retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (pp. 11-19).
- Zhu, Y., Duan, H., Wang, Z., Kim, E. H., Fu, Z., & Pedrycz, W. (2025). Robust Classification via Interval Type-2 Fuzzy C-Means and Gradient Boosting. *IEEE Transactions on Fuzzy Systems*, 33(9), 3103-3117.
- Huang, Y., Zhang, C., & Pan, C. (2022). Channel-aided transmission parameter signalling detection for DTMB-A. *IEEE Transactions on Broadcasting*, 69(1), 303-312.