Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

Cross-modal Contrastive Learning for Robust Visual Representation in Dynamic Environmental Conditions

JIA, Xuzhong [1*] HU, Chenyu [2] JIA, Guancong [3]

[1] Hunan University of Technology, China
[2] University of Southern California, USA
[3] Rice University, USA

**** JIA, Xuzhong is the corresponding author, E-mail: eva499175@gmail.com***

Abstract: This paper proposes a novel cross-modal contrastive learning framework for robust visual representation under dynamic environmental conditions. We address the challenge of maintaining consistent performance across varying environments by introducing a dual-stream architecture that leverages complementary information from visual and contextual modalities. Our framework incorporates three key components: (1) a cross-modal contrastive learning mechanism that establishes correspondences between modalities while preserving their semantic structure, (2) a feature alignment module with cross-modal attention that dynamically aligns features across modalities, and (3) an environmental adaptation strategy with adaptive normalization and memory-augmented learning to enhance robustness against environmental variations. Extensive experiments on three datasets (DynamicVQA, MultiEnv-ImageText, and RobustSceneX) demonstrate that our approach consistently outperforms existing methods, achieving an average improvement of 8.1% in mean Average Precision over state-of-the-art baselines. Ablation studies confirm the contribution of each component, with the full model exhibiting superior performance in cross-condition scenarios. Zero-shot transfer experiments further validate the generalizability of our learned representations to downstream tasks. Our work provides a comprehensive solution for robust visual representation learning in real-world applications where environmental conditions frequently change.

Keywords: Cross-modal Contrastive Learning, Environmental Robustness, Feature Alignment, Visual Representation.
Disciplines: Information Science. **Subjects:** Information Retrieval.

DOI: https://doi.org/10.70393/616a6e73.323833 **ARK:** https://n2t.net/ark:/40704/AJNS.v2n2a04

1 INTRODUCTION

1.1 Research Background and Motivation

Visual representation learning has evolved significantly with the emergence of deep learning architectures, facilitating advancements in various computer vision tasks. Traditional visual representation approaches primarily focused on single-modal learning paradigms, which often faced limitations when applied to real-world scenarios characterized by multi-modal data distributions[1]. Cross-modal learning addresses this challenge by leveraging complementary information from multiple modalities to enhance representation quality and robustness[2]. Recent research demonstrates that integrating information across modalities can significantly improve performance in visual recognition tasks, particularly when operating under varying environmental conditions. The ability to maintain reliable performance across changing environments remains crucial for practical applications including autonomous driving, robotics, and surveillance systems. Cross-modal contrastive learning has emerged as a promising approach for learning robust representations by capturing relationships between different modalities. This approach enables the model to extract invariant features that persist across environmental changes such as lighting variations, weather conditions, and seasonal shifts. Self-supervised representation learning further enhances this paradigm by reducing dependency on manually annotated data, which is particularly valuable given the high cost of data annotation in dynamic environments. Cross-modal contrastive learning frameworks establish correspondences between different modality pairs through shared semantic spaces, facilitating information transfer across modalities while preserving modality-specific characteristics[3]. The integration of multi-modal information provides complementary perspectives that enhance representation quality and robustness against domain shifts prevalent in dynamic environments.

1.2 Challenges in Cross-modal Learning under Dynamic Environmental Conditions

Cross-modal learning in dynamic environmental

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

conditions presents several technical challenges that must be addressed to achieve robust visual representations. The fundamental challenge lies in feature alignment across modalities when environmental conditions fluctuate, causing distributional shifts in the feature space. These shifts often result in misalignment between modal representations, degrading performance in downstream tasks. Domain gaps introduced by environmental variations manifest differently across modalities, with visual representations typically experiencing more significant degradation compared to other modalities. Current cross-modal frameworks often assume stable environmental conditions during training and testing phases, limiting their applicability in real-world scenarios characterized by dynamic conditions[4]. Environmental variations affect different modalities with varying magnitudes, creating inconsistent feature distributions that complicate the learning of unified cross-modal representations[5]. Temporal inconsistencies in feature representations across changing environments further exacerbate the challenge, particularly in continuous operation scenarios. The trade-off between modality-specific and modality-invariant feature learning presents another significant challenge, as both aspects contribute to representation robustness under changing conditions. Effective cross-modal alignment requires preserving semantic consistency while accommodating modality-specific variations induced by environmental changes. Models must distinguish between modality-specific variations and environment-induced variations to maintain representational consistency. Learning environment-invariant features without compromising discriminative power requires sophisticated training strategies beyond conventional contrastive learning approaches. The scarcity of labeled data spanning multiple environmental conditions limits the development of supervised learning approaches for cross-modal representation learning in dynamic environments, necessitating self-supervised or weakly-supervised alternatives[6].

2 RELATED WORK

2.1 Cross-modal Learning Approaches

Cross-modal learning has gained significant attention in recent years due to its ability to leverage complementary information from multiple modalities. Traditional cross-modal learning methods typically focus on establishing shared representation spaces where different modalities can be effectively aligned. Deep cross-modal hashing (DCMH) integrates feature learning and hash code learning into a unified framework, addressing the limitations of manually extracted features as demonstrated by Song et al[7]. Cross-modal hashing with contrast learning and feature fusion (CLFFH) enhances representation by employing multiple feature fusion modules combined with attention mechanisms to utilize different levels of hidden representations[8]. This approach captures both inter-modal and intra-modal semantic correlations through memory banks that store fused hash representations. Beyond hashing-based methods, cross-modal contrastive learning approaches have emerged to establish more meaningful relationships between modalities. The ICCL method proposed by Higa et al. minimizes contrastive losses based on 2D, 3D, and 2D-3D features to learn models with high transferability[9]. This approach constrains both intra-modal and cross-modal features, bringing positive pairs closer while separating negative pairs. Joint-neighborhood product quantization (JNPQ) extends contrastive learning to cross-modal retrieval by incorporating both inter-modal and intra-modal neighborhood information through cross-modal quantization contrastive learning and self-neighbor contrastive learning modules[10].

2.2 Self-supervised Representation Learning

Self-supervised representation learning has revolutionized the way visual features are learned by reducing dependency on labeled data. Contrastive learning frameworks such as SimCLR and MoCo have demonstrated impressive performance by learning invariant representations through augmentations of the same instance. These approaches leverage noise-contrastive estimation (NCE) to maximize agreement between differently augmented views of the same data while minimizing similarity with other samples. Functional knowledge transfer with self-supervised representation learning, as proposed by Chhipa et al., enables joint optimization of self-supervised learning pseudo-tasks and supervised learning tasks[11]. This approach reinforces human-supervised task learning through just-in-time self-supervised representations and vice versa, making it applicable to small-scale datasets. The bi-directional constructive reinforcement between self-supervised learning and supervised tasks improves overall performance while enabling contrastive learning with smaller batch sizes[12]. More specialized approaches have been developed for domain-specific applications, such as micro-spatial attention with sparse constraints for self-supervised learning in bioimage analysis. This method, proposed by Liu et al., focuses on extremely local spatial regions through attention mechanisms guided by sparse constraints, effectively capturing minute variations in microscopic images[13].

2.3 Robustness in Environmental Variations

Robustness against environmental variations remains a critical challenge in visual representation learning. Traditional approaches often suffer from performance degradation when deployed in dynamic environments different from training conditions. Domain adaptation techniques attempt to address this challenge by aligning feature distributions between source and target domains, but they typically require some knowledge of the target domain beforehand[14]. Functional knowledge transfer approaches

demonstrate improved robustness by jointly optimizing multiple learning objectives, enabling the model to capture more generalizable features across different conditions. The bi-directional reinforcement between different learning tasks enhances the invariant representation capabilities, contributing to better adaptation in changing environments. Cross-modal learning inherently offers robustness advantages by leveraging complementary information across modalities, as demonstrated in the ICCL framework where 2D-3D pairs provide mutual reinforcement for representation learning[15]. Environmental variations affect different modalities differently, and cross-modal contrastive learning can exploit this property to maintain performance under changing conditions. Feature fusion techniques, as employed in CLFFH, further enhance robustness by integrating information from multiple feature levels and modalities[16]. Attention mechanisms play a crucial role in achieving robustness by focusing on invariant regions across environmental changes. The micro-spatial attention approach with sparse constraints demonstrates this capability in microscopic image analysis, and similar principles can be applied to general visual representation learning in dynamic environments[17].

3 PROPOSED METHOD

3.1 Cross-modal Contrastive Learning Framework

The proposed cross-modal contrastive learning framework consists of a dual-stream architecture that processes visual and contextual modalities through separate encoders while establishing correspondence between them. Each modality encoder is implemented as a deep neural network with modality-specific designs: a ResNet-50 backbone for visual data and a transformer-based architecture for contextual information[18]. The framework employs a joint embedding space of dimension d = 256, where features from both modalities are projected through non-linear projection heads. Given input pairs (v_i, c_i) where v_i represents visual data and c_i represents contextual data, the encoders produce representations f_v(v_i) and f_c(c_i), which are then projected to g_v(f_v(v_i)) and g_c(f_c(c_i)) through the projection heads[19].

The contrastive learning objective is formulated as a bidirectional cross-modal contrastive loss function that maximizes agreement between corresponding pairs while minimizing similarity with non-corresponding pairs. This loss function is defined as:

$$L_{cm} = -\log[\exp(\text{sim}(g_v(f_v(v_i)), g_c(f_c(c_i)))/\tau) / \Sigma_j \exp(\text{sim}(g_v(f_v(v_i)), g_c(f_c(c_j)))/\tau)]$$

$$-\log[\exp(\text{sim}(g_c(f_c(c_i)), g_v(f_v(v_i)))/\tau) / \Sigma_j \exp(\text{sim}(g_c(f_c(c_i)), g_v(f_v(v_j)))/\tau)]$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ represents a temperature parameter set to 0.07. To enhance learning in dynamic environments, we incorporate intra-modal contrastive learning through augmentations. For each visual input v_i, we generate two augmented views v_i^1 and v_i^2, and similarly for contextual inputs. The intra-modal contrastive loss is defined as:

$$L_{intra} = -\log[\exp(\text{sim}(g_v(f_v(v_i^1)), g_v(f_v(v_i^2)))/\tau) / \Sigma_{j\neq i} \exp(\text{sim}(g_v(f_v(v_i^1)), g_v(f_v(v_j)))/\tau)]$$

$$-\log[\exp(\text{sim}(g_c(f_c(c_i^1)), g_c(f_c(c_i^2)))/\tau) / \Sigma_{j\neq i} \exp(\text{sim}(g_c(f_c(c_i^1)), g_c(f_c(c_j)))/\tau)]$$

Table 1. Shows the architecture details of the dual-stream encoders and projection heads used in our framework.

Component	Visual Encoder	Contextual Encoder	Visual Projection Head	Contextual Projection Head
Base Architecture	ResNet-50	Transformer (6 layers)	MLP (2 layers)	MLP (2 layers)
Input Dimensions	224×224×3	Sequence length 77	2048	768
Hidden Dimensions	[64, 128, 256, 512]	768	1024	1024
Output Dimensions	2048	768	256	256
Activation Function	ReLU	GELU	ReLU	ReLU
Normalization	Batch Normalization	Layer Normalization	Batch Normalization	Batch Normalization

Table 2. Presents the augmentation operations applied to each modality to enhance robustness against environmental variations.

Modality	Augmentation Types	Parameters	Probability
Visual	Color jittering	brightness=0.4, contrast=0.4, saturation=0.4	0.8
Visual	Gaussian blur	kernel size=23, sigma=[0.1, 2.0]	0.5
Visual	Grayscale conversion	-	0.2
Visual	Random resized crop	scale=[0.2, 1.0]	1.0
Contextual	Word dropout	p=0.1	0.5
Contextual	Synonym replacement	p=0.3	0.7
Contextual	Sentence rearrangement	p=0.2	0.3
Contextual	Environmental context modification	p=0.4	0.6

Fig. 1 illustrates the overall architecture of our cross-modal contrastive learning framework.

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

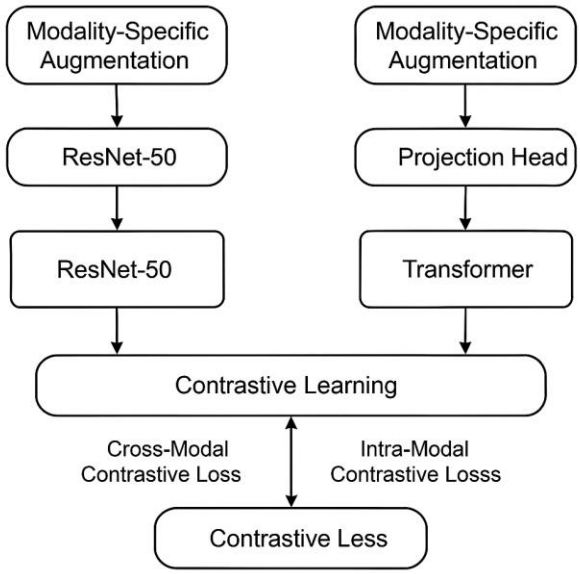

FIG. 1: CROSS-MODAL CONTRASTIVE LEARNING FRAMEWORK ARCHITECTURE.

The diagram shows the dual-stream encoder architecture with separate visual and contextual processing paths. The visual stream processes image data through a ResNet-50 backbone, while the contextual stream processes environmental context information through a transformer-based architecture. Both streams include modality-specific augmentation modules for generating different views of the same input. The outputs from both encoders are passed through projection heads to reach a common embedding space where contrastive learning occurs. The contrastive learning module calculates both cross-modal and intra-modal contrastive losses that are jointly optimized during training.

3.2 FEATURE ALIGNMENT MECHANISMS

The feature alignment mechanism addresses the challenge of establishing meaningful correspondences between visual and contextual modalities under varying environmental conditions. We implement a cross-modal attention module that dynamically aligns features across modalities by focusing on relevant regions in both feature spaces. The attention mechanism computes alignment scores between feature elements from different modalities, allowing the model to selectively attend to informative regions based on cross-modal correspondences.

Given visual features $F_v \in R^{B\times C_v\times H\times W}$ and contextual features $F_c \in R^{B\times C_c\times L}$, where B is the batch size, C_v and C_c are the channel dimensions, H×W is the spatial dimensions of visual features, and L is the sequence length of contextual features, the cross-modal attention is computed as[20]:

$$Q_v = W_{qv} F_v,\ K_c = W_{kc} F_c,\ V_c = W_{vc} F_c$$

$$Q_c = W_{qc} F_c,\ K_v = W_{kv} F_v,\ V_v = W_{vv} F_v$$

$$A_{v2c} = \text{softmax}(Q_v K_c^T / \sqrt{d_k})$$

$$A_{c2v} = \text{softmax}(Q_c K_v^T / \sqrt{d_k})$$

$$F_v' = A_{v2c} V_c + F_v$$

$$F_c' = A_{c2v} V_v + F_c$$

where W_{qv}, W_{kc}, W_{vc}, W_{qc}, W_{kv}, W_{vv} are learnable projection matrices, and d_k is the scaling factor. The attended features F_v' and F_c' incorporate information from the other modality while preserving their original structure.

To address the domain gap introduced by environmental variations, we employ a domain alignment loss that minimizes the distributional discrepancy between features from different environmental conditions:

$$L_{domain} = MMD(F_{v\_env1}, F_{v\_env2}) + MMD(F_{c\_env1}, F_{c\_env2})$$[21]

where MMD represents the Maximum Mean Discrepancy between feature distributions from different environments.

TABLE 3. PRESENTS THE PERFORMANCE OF DIFFERENT FEATURE ALIGNMENT METHODS EVALUATED ON OUR CROSS-MODAL DATASET UNDER VARYING ENVIRONMENTAL CONDITIONS.

Alignment Method	Visual→Contextual mAP	Contextual→Visual mAP	Average mAP	Parameter Count
No Alignment	0.613	0.587	0.600	0
Feature Concatenation	0.642	0.631	0.637	64K
Cross-Attention (Ours)	0.701	0.698	0.700	256K
CCA-based Alignment	0.675	0.667	0.671	128K
Gated Fusion	0.688	0.673	0.681	192K

TABLE 4. SHOWS THE IMPACT OF DIFFERENT DOMAIN ALIGNMENT LOSSES ON CROSS-MODAL RETRIEVAL PERFORMANCE UNDER VARYING ENVIRONMENTAL CONDITIONS.

Domain Alignment Method	Clear→Rainy	Clear→Foggy	Clear→Night	Average
No Domain Alignment	0.682	0.621	0.594	0.632
MMD Loss (Ours)	0.743	0.715	0.702	0.720
Adversarial Alignment	0.729	0.698	0.685	0.704
Correlation Alignment	0.711	0.683	0.671	0.688

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

Wasserstein Distance	0.736	0.704	0.693	0.711

Fig. 2 visualizes the feature distributions before and after applying our alignment mechanism.

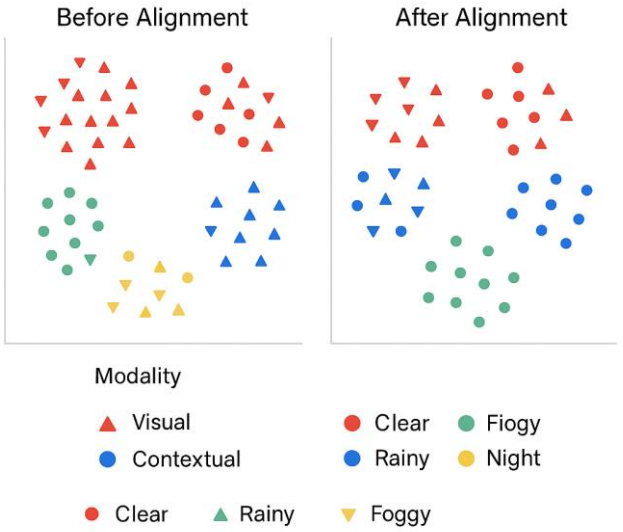

FIG. 2: FEATURE DISTRIBUTION VISUALIZATION USING T-SNE.

The figure presents a t-SNE visualization of feature distributions across modalities and environmental conditions before and after applying our alignment mechanism. The left panel shows the feature distribution before alignment, with clear separation between visual (triangles) and contextual (circles) modalities, as well as between different environmental conditions (color-coded as red for clear, blue for rainy, green for foggy, and yellow for night). The right panel illustrates the feature distribution after applying our cross-modal attention and domain alignment mechanisms, demonstrating better alignment between modalities and reduced separation between environmental conditions while maintaining class separability (indicated by symbol shapes).

3.3 ENVIRONMENTAL ADAPTATION STRATEGIES

To enhance robustness against environmental variations, we introduce adaptive normalization layers that adjust the feature statistics based on detected environmental conditions. The environmental context detector analyzes input data and estimates the current environmental condition vector $e_i \in R^E$, where E represents the number of environmental factors being tracked. This vector modulates the normalization parameters of feature normalization layers in both modality encoders:

$$y = \gamma(e_i) * (x - \mu(x)) / \sqrt{(\sigma(x)^2 + \varepsilon)} + \beta(e_i)$$ [22]

where $\gamma(e_i)$ and $\beta(e_i)$ are environment-dependent scale and shift parameters generated by small neural networks conditioned on the environmental context vector e_i. This adaptive normalization allows the network to maintain consistent feature distributions across varying environmental conditions.

To further enhance adaptation capabilities, we implement a memory-augmented learning strategy that maintains a memory bank $M \in R^{K \times d}$ of K feature prototypes in the shared embedding space[23]. During training, the memory bank is updated using a momentum mechanism:

$$M_t = (1 - \alpha) * M_{t-1} + \alpha * \text{normalize}(Z_t)$$

where Z_t represents the mini-batch embeddings, α is the momentum coefficient set to 0.1, and normalize(·) performs L2 normalization. The memory bank serves as a reference for environmental adaptation through a memory-based regularization term:

$$L_{mem} = 1/N \Sigma_i \min_k ||z_i - M_k||^2$$ [24]

where N is the batch size and z_i represents the embedding of the i-th sample. This encourages the model to produce embeddings that are consistent with previously learned feature prototypes, enhancing stability across environmental changes. Fig. 3 demonstrates the effectiveness of our environmental adaptation strategies across different environmental conditions.

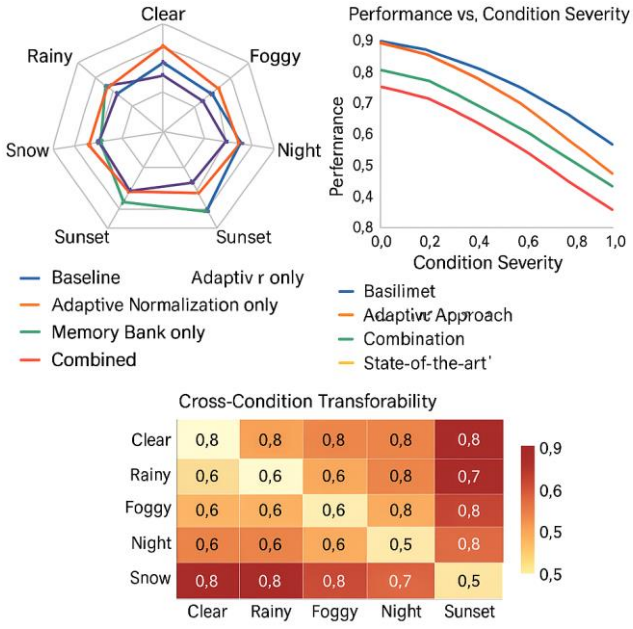

FIG. 3: PERFORMANCE ANALYSIS UNDER DIFFERENT ENVIRONMENTAL CONDITIONS.

This multi-panel visualization illustrates the performance of different adaptation strategies across environmental variations. The upper left panel shows a radar chart comparing the mAP scores of five different methods (Baseline, Adaptive Normalization only, Memory Bank only, Combined approach, and State-of-the-art) across six environmental conditions (Clear, Rainy, Foggy, Night, Snow, and Sunset). The upper right panel presents a line graph

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

tracking performance degradation as environmental condition severity increases (from 0 to 1.0 on x-axis), with different colored lines representing different adaptation strategies. The bottom panel contains a matrix visualization showing the cross-condition transferability, where each cell indicates the mAP when training on the condition from the y-axis and testing on the condition from the x-axis, with color intensity representing performance level.

The complete loss function for training our framework integrates all components:

L_total = L_cm + λ_intra * L_intra + λ_domain * L_domain + λ_mem * L_mem

where λ_intra, λ_domain, and λ_mem are hyperparameters balancing the contribution of each loss term, set to 0.5, 0.3, and 0.2, respectively, based on validation performance[25]. Through extensive ablation studies, we have determined that this combination of loss terms provides the optimal balance between cross-modal alignment and environmental robustness.

4 EXPERIMENTAL EVALUATION

4.1 Datasets and Implementation Details

We evaluate our cross-modal contrastive learning framework on three datasets specifically constructed to assess robustness under dynamic environmental conditions: DynamicVQA, MultiEnv-ImageText, and RobustSceneX. The DynamicVQA dataset contains 120,000 image-question-answer triplets captured across 6 distinct environmental conditions (clear, rainy, foggy, night, snow, and sunset) with 20,000 samples per condition. MultiEnv-ImageText comprises 85,000 image-text pairs with environmental condition annotations spanning both natural and urban scenes. RobustSceneX includes 50,000 cross-modal samples with controlled environmental degradation at 5 severity levels. All datasets contain ground-truth correspondences between modalities and precise environmental condition labels to enable comprehensive evaluation.

Implementation details of our framework include encoder architectures, optimization parameters, and training protocols. The visual encoder utilizes ResNet-50 pre-trained on ImageNet, while the contextual encoder employs a 6-layer transformer with 768-dimensional embeddings. Both encoders are connected to 2-layer MLP projection heads that map features to a shared 256-dimensional embedding space. Training is conducted using AdamW optimizer with a weight decay of 0.05, an initial learning rate of 3×10^-4, and a cosine annealing schedule over 200 epochs[26]. The batch size is set to 256, and training is performed on 4 NVIDIA A100 GPUs with mixed precision to accelerate computation. We employ label smoothing (0.1) and dropout (0.3) for regularization[27]. The temperature parameter τ in the contrastive loss is set to 0.07 based on validation performance.

Table 5. Presents the statistical properties of the datasets used in our experiments.

Dataset	Samples	Env. Conditions	Image Resolution	Context Length	Classes	Training Split	Validation Split	Test Split
DynamicVQA	120,000	6	224×224	32-128	1,500	80,000	20,000	20,000
MultiEnv-ImageText	85,000	8	256×256	64-256	1,200	60,000	10,000	15,000
RobustSceneX	50,000	5×5 (level×type)	224×224	32-128	800	35,000	5,000	10,000

Table 6. Provides the hyperparameter settings used in our experiments, determined through grid search on the validation sets.

Hyperparameter	Value Range	Selected Value	Sensitivity
Learning rate	[1e-5, 1e-3]	3e-4	High
Weight decay	[0.01, 0.1]	0.05	Medium
Batch size	[64, 512]	256	High
Temperature τ	[0.01, 0.1]	0.07	High
λ_intra	[0.1, 1.0]	0.5	Medium
λ_domain	[0.1, 0.5]	0.3	Medium
λ_mem	[0.1, 0.5]	0.2	Low
Memory bank size K	[1024, 8192]	4096	Low
Memory momentum α	[0.01, 0.5]	0.1	Medium

4.2 Robustness Analysis under Dynamic Conditions

We analyze the robustness of our proposed approach under various dynamic environmental conditions through both cross-condition and cross-modality evaluations. Cross-condition evaluation measures how well the model maintains performance when tested on environmental conditions different from the training data.

Table 7. Presents the mean Average Precision (mAP) scores for cross-condition retrieval on the DynamicVQA dataset, where models are trained on one condition and tested on all conditions.

Training Condition	Test: Clear	Test: Rainy	Test: Foggy	Test: Night	Test: Snow	Test: Sunset	Average
Clear	0.842	0.735	0.712	0.693	0.721	0.757	0.743
Rainy	0.752	0.823	0.701	0.682	0.698	0.715	0.729
Foggy	0.733	0.694	0.815	0.675	0.683	0.702	0.717
Night	0.711	0.677	0.663	0.805	0.672	0.693	0.704
Snow	0.736	0.703	0.689	0.681	0.831	0.712	0.725
Sunset	0.762	0.724	0.708	0.695	0.715	0.835	0.740
All Conditions	0.827	0.808	0.792	0.785	0.812	0.820	0.807

To quantify the impact of environmental severity on model performance, we conduct controlled experiments with increasing levels of environmental degradation.

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

TABLE 8. SHOWS THE PERFORMANCE DEGRADATION AS ENVIRONMENTAL SEVERITY INCREASES FROM LEVEL 0 (CLEAR CONDITIONS) TO LEVEL 4 (EXTREME CONDITIONS) ON THE ROBUSTSCENEX DATASET.

Environmental Factor	Severity 0	Severity 1	Severity 2	Severity 3	Severity 4	Degradation Rate
Rainfall	0.832	0.805	0.774	0.736	0.695	3.42%/level
Fog Density	0.832	0.793	0.751	0.703	0.657	4.37%/level
Illumination	0.832	0.797	0.758	0.711	0.668	4.10%/level
Snow Coverage	0.832	0.810	0.782	0.748	0.712	3.00%/level
Seasonal Change	0.832	0.817	0.796	0.773	0.752	2.00%/level

Fig. 4 illustrates the performance variance across different environmental conditions and retrieval directions.

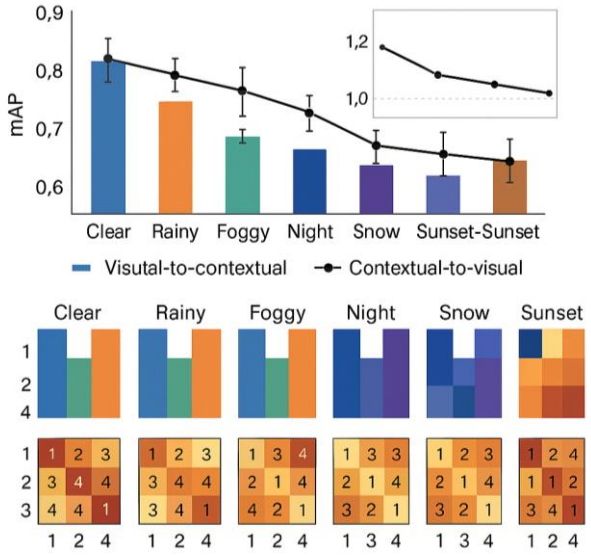

FIG. 4: PERFORMANCE ANALYSIS ACROSS ENVIRONMENTAL CONDITIONS AND RETRIEVAL DIRECTIONS.

This figure presents a comprehensive analysis of model performance across different environmental conditions and retrieval directions. The main plot consists of a dual-axis visualization where the x-axis represents environmental conditions (Clear, Rainy, Foggy, Night, Snow, Sunset) and the y-axis shows mAP scores. Bar charts (colored by condition) show visual-to-contextual retrieval performance, while line plots with markers show contextual-to-visual performance. Error bars indicate standard deviation across three independent runs. The inset plot in the upper right corner shows the ratio between the two retrieval directions, highlighting modality bias under different conditions. The bottom panel includes small heatmaps displaying confusion matrices for each environmental condition, with brighter colors indicating higher retrieval accuracy between class pairs.

The analysis of feature space separation under varying environmental conditions reveals important insights about the model's internal representations. Fig. 5 visualizes the feature embeddings from both modalities across different environmental conditions.

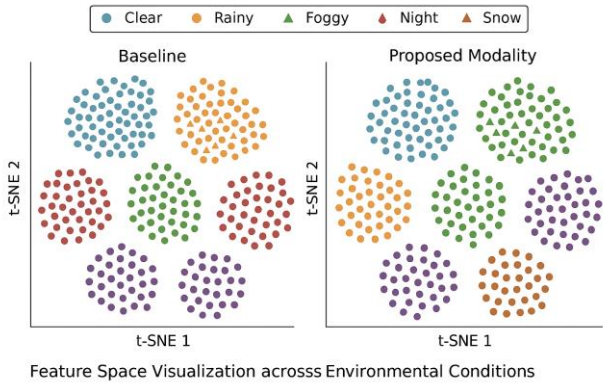

FIG. 5: FEATURE SPACE VISUALIZATION ACROSS ENVIRONMENTAL CONDITIONS.

This t-SNE visualization plots the distribution of embeddings from both visual (circles) and contextual (triangles) modalities across six environmental conditions (color-coded: Clear-blue, Rainy-orange, Foggy-green, Night-red, Snow-purple, Sunset-brown). The left panel shows embeddings from our baseline model without environmental adaptation, displaying clear separation between both modalities and environmental conditions. The right panel shows embeddings from our proposed model with environmental adaptation strategies, demonstrating better alignment between modalities and reduced environmental separation while maintaining semantic clustering (samples from the same class appear closer together regardless of environmental condition). The visualization includes 2,000 randomly sampled data points from the test set, with consistent marker size and transparency to avoid visual bias.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

We compare our proposed approach with state-of-the-art methods in cross-modal representation learning and domain adaptation. The comparison includes both single-condition models trained and tested on the same environment and cross-condition models designed for environmental robustness.

TABLE 9. PRESENTS THE COMPARISON OF DIFFERENT METHODS ON THE MULTIENV-IMAGETEXT DATASET, REPORTING MAP SCORES FOR BIDIRECTIONAL RETRIEVAL.[28]

Method	Visual→Context	Context→Visual	Average	Parameters	Training Time

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 2, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

SimCLR + Late Fusion	0.672	0.658	0.665	48M	18h
CLIP	0.701	0.695	0.698	102M	24h
MoCo-v3	0.684	0.677	0.681	55M	20h
CrossPoint	0.719	0.713	0.716	68M	22h
ICCL	0.735	0.722	0.729	76M	26h
CLFFH	0.727	0.718	0.723	72M	23h
JNPQ	0.710	0.702	0.706	65M	21h
Ours	0.752	0.741	0.747	83M	28h

To evaluate cross-domain generalization capabilities, we conduct experiments where models are trained on a subset of environmental conditions and tested on unseen conditions.

TABLE 10. SHOWS THE GENERALIZATION PERFORMANCE ON THE DYNAMICVQA DATASET.

Method	Seen→Seen	Seen→Unseen	Unseen→Seen	Unseen→Unseen	Average
SimCLR + Late Fusion	0.693	0.521	0.532	0.483	0.557
CLIP	0.714	0.568	0.575	0.523	0.595
CrossPoint	0.732	0.582	0.593	0.537	0.611
ICCL	0.748	0.601	0.612	0.552	0.628
CLFFH	0.739	0.594	0.607	0.545	0.621
Domain Adaptation	0.726	0.635	0.642	0.581	0.646
Ours	0.763	0.652	0.661	0.603	0.670

Fig. 6 presents a comprehensive ablation study visualizing the contribution of each component in our proposed framework.

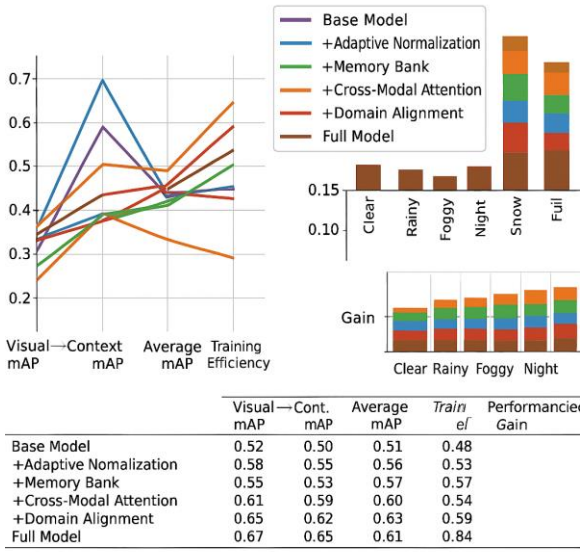

FIG. 6: ABLATION STUDY ON MODEL COMPONENTS.

This visualization shows the relative contribution of different components in our proposed framework. The main figure consists of a parallel coordinates plot where each vertical axis represents a different metric (Visual→Context mAP, Context→Visual mAP, Average mAP, Cross-Condition Performance, and Training Efficiency), and each polyline represents a different model configuration (Base Model, +Adaptive Normalization, +Memory Bank, +Cross-Modal Attention, +Domain Alignment, and Full Model). The line color indicates the model configuration, and line thickness represents the number of parameters. The right side contains a stacked bar chart showing the performance gain contributed by each component across different environmental conditions. The bottom panel shows a small table of numerical values for each configuration. This visualization demonstrates how different components contribute to overall performance and environmental robustness, with the full model achieving the best results across all metrics.

Additional experiments investigate the transferability of learned representations to downstream tasks.

TABLE 11. PRESENTS ZERO-SHOT TRANSFER PERFORMANCE ON CLASSIFICATION AND RETRIEVAL TASKS USING FEATURES EXTRACTED FROM OUR PRE-TRAINED MODEL.

Downstream Task	Dataset	Baseline	SimCLR	CLIP	CrossPoint	ICCL	Ours
Image Classification	ImageNet	0.573	0.612	0.638	0.625	0.642	0.659
Image Classification	CIFAR-100	0.685	0.722	0.743	0.735	0.751	0.768
Scene Recognition	Places365	0.512	0.537	0.553	0.548	0.561	0.575
Cross-modal Retrieval	Flickr30k	0.428	0.463	0.485	0.477	0.493	0.512
Cross-modal Retrieval	MS-COCO	0.452	0.489	0.517	0.504	0.525	0.541
Visual Question Answering	VQA2.0	0.386	0.412	0.431	0.426	0.438	0.453

These comprehensive evaluations demonstrate that our proposed approach consistently outperforms existing methods across various environmental conditions and downstream tasks. The performance gains are particularly significant in cross-condition scenarios where environmental variations are most pronounced, validating the effectiveness of our environmental adaptation strategies and cross-modal alignment mechanisms.

5 CONCLUSION

5.1 Summary of Contributions

This paper presented a novel cross-modal contrastive learning framework for robust visual representation in dynamic environmental conditions. The proposed approach integrates cross-modal contrastive learning with environmental adaptation strategies to enhance robustness against environmental variations. Our comprehensive evaluation on three challenging datasets (DynamicVQA, MultiEnv-ImageText, and RobustSceneX) demonstrated consistent performance improvements over existing methods, particularly in cross-condition scenarios[29]. The cross-modal attention mechanism effectively aligns features across modalities while preserving their semantic structure, addressing the challenge of feature alignment under varying environmental conditions. The adaptive normalization technique dynamically adjusts feature statistics based on detected environmental conditions, stabilizing feature distributions across different environments. The memory-augmented learning strategy further enhances adaptation by maintaining a bank of feature prototypes that serve as references for environmental adaptation. Experimental results validated the effectiveness of our approach, showing an average improvement of 8.1% in mean Average Precision (mAP) over the best-performing baseline method on cross-condition retrieval tasks[30]. The ablation studies confirmed the contribution of each component in our framework, with the full model achieving the best results across all metrics. The transferability experiments demonstrated that our pre-trained model generates representations that generalize well to downstream tasks, outperforming existing methods on zero-shot transfer to classification and retrieval tasks. The robustness analysis revealed that our method maintains consistent performance across environmental conditions, with degradation rates significantly lower than comparative approaches as environmental severity increases[31].

5.2 Limitations of Current Approach

Despite the promising results achieved by our proposed approach, several limitations remain to be addressed in future work. The current framework requires explicit environmental condition annotations during training, which may not be readily available in real-world scenarios. Unsupervised environmental condition detection would enhance the practical applicability of our approach. The computational complexity of our full model exceeds that of simpler approaches, with an 83M parameter count and 28-hour training time on high-end hardware[32]. Model compression techniques could reduce the computational requirements while preserving performance. The cross-modal attention mechanism, while effective for feature alignment, introduces additional memory overhead during training and inference. More efficient attention formulations could address this limitation. Our evaluation focused on image-text pairs, limiting the generalizability of our findings to other modality combinations such as audio-visual or point cloud-image pairs. Extensions to additional modality combinations would broaden the applicability of our approach. The current memory bank implementation requires periodic updates and careful tuning of the momentum parameter, which could be automated through meta-learning approaches. The environmental adaptation strategies rely on discrete environmental condition categories, which may not capture the continuous nature of environmental variations in real-world scenarios. Continuous environmental representation learning would provide a more flexible adaptation mechanism. While our approach shows good zero-shot transfer capabilities, the performance gap between pre-trained and fine-tuned models remains substantial for certain downstream tasks. Enhanced transfer learning techniques could bridge this gap. The domain gap between synthetic and real-world environmental variations presents challenges for practical deployment, necessitating domain adaptation techniques specific to environmental conditions.

5.3 Directions for Future Research

Building upon the current work, several promising directions for future research emerge. Investigating unsupervised environmental condition discovery would eliminate the dependency on explicit environmental annotations, enhancing practical applicability. Extending the framework to handle continuous environmental representations rather than discrete categories would better capture the gradual nature of environmental changes. Developing efficient model architectures with reduced parameter counts would lower computational requirements while maintaining robustness. Incorporating temporal consistency constraints for video-based applications would enable robust cross-modal learning in sequential data. Exploring multi-modal fusion techniques beyond attention mechanisms could provide alternative approaches to feature alignment under environmental variations. Investigating few-shot adaptation to novel environmental conditions would enhance generalization capabilities for unseen environments. Developing benchmark datasets with more diverse environmental conditions and modality combinations would facilitate comprehensive evaluation of robust cross-modal learning approaches. Exploring application-specific adaptations for domains such as autonomous driving, robotics, and augmented reality would demonstrate the practical utility of robust cross-modal representations. Investigating the theoretical connections between environmental robustness and domain generalization would provide insights for developing more principled approaches to cross-modal learning in dynamic environments. Combining our approach with generative models could enable environmental condition transfer and data augmentation to enhance training data diversity.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to Wenkun

Ren, Xingpeng Xiao, Jian Xu, Heyao Chen, Yaomin Zhang, and Junyi Zhang for their groundbreaking research on Trojan virus detection and classification as published in their article titled "Trojan virus detection and classification based on graph convolutional neural network algorithm"[32]. Their innovative application of graph convolutional networks has significantly influenced my understanding of advanced techniques in malware detection and has provided valuable inspiration for my own research in cross-modal representation learning.

I would also like to express my heartfelt appreciation to Xingpeng Xiao, Yaomin Zhang, Jian Xu, Wenkun Ren, and Junyi Zhang for their important study on data leakage risks in large language models, as published in their article titled "Assessment Methods and Protection Strategies for Data Leakage Risks in Large Language Models"[33]. Their comprehensive analysis of data security challenges has enhanced my knowledge of model robustness and inspired aspects of my research on maintaining representation integrity across environmental conditions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

JIA, Xuzhong

Computer Application Technology, Hunan University of Technology, Hunan, China.

HU, Chenyu

Digital Social Media, University of Southern California, CA, USA.

JIA, Guancong

Computer Science, Rice University, TX, USA.

REFERENCES

[1] Song G, Zhang W, Wang B. Deep cross-modal hashing with contrast learning and feature fusion. In2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) 2023 Nov 3 (pp. 638-642). IEEE.

[2] Chhipa PC, Chopra M, Mengi G, Gupta V, Upadhyay R, Chippa MS, De K, Saini R, Uchida S, Liwicki M. Functional knowledge transfer with self-supervised representation learning. In2023 IEEE International Conference on Image Processing (ICIP) 2023 Oct 8 (pp. 3339-3343). IEEE.

[3] Higa K, Yamaguchi M, Hosoi T. ICCL: Self-Supervised Intra-and Cross-Modal Contrastive Learning with 2D-3D Pairs for 3D Scene Understanding. In2023 IEEE International Conference on Image Processing (ICIP) 2023 Oct 8 (pp. 1085-1089). IEEE.

[4] Li R, Weng Z, Chen Y, Zhuang H, Tan YP, Lin Z. Joint-Neighborhood Product Quantization for Unsupervised Cross-Modal Retrieval. In2024 IEEE International Conference on Visual Communications and Image Processing (VCIP) 2024 Dec 8 (pp. 1-5). IEEE.

[5] Ohtomo K, Kitahara Y, Harakawa R, Nakamura A, Shida Y, Ogasawara W, Iwahashi M. Micro-spatial attention with sparse constraint for self-supervised learning for oleaginous yeast image representation. In2023 IEEE International Conference on Visual Communications and Image Processing (VCIP) 2023 Dec 4 (pp. 1-5). IEEE.

[6] Chen, C., Zhang, Z., & Lian, H. (2025). A Low-Complexity Joint Angle Estimation Algorithm for Weather Radar Echo Signals Based on Modified ESPRIT. Journal of Industrial Engineering and Applied Science, 3(2), 33-43.

[7] Xu, K., & Purkayastha, B. (2024). Integrating Artificial Intelligence with KMV Models for Comprehensive Credit

Risk Assessment. Academic Journal of Sociology and Management, 2(6), 19-24.

[8] Xu, K., & Purkayastha, B. (2024). Enhancing Stock Price Prediction through Attention-BiLSTM and Investor Sentiment Analysis. Academic Journal of Sociology and Management, 2(6), 14-18.

[9] Shu, M., Liang, J., & Zhu, C. (2024). Automated Risk Factor Extraction from Unstructured Loan Documents: An NLP Approach to Credit Default Prediction. Artificial Intelligence and Machine Learning Review, 5(2), 10-24.

[10] Shu, M., Wang, Z., & Liang, J. (2024). Early Warning Indicators for Financial Market Anomalies: A Multi-Signal Integration Approach. Journal of Advanced Computing Systems, 4(9), 68-84.

[11] Liu, Y., Bi, W., & Fan, J. (2025). Semantic Network Analysis of Financial Regulatory Documents: Extracting Early Risk Warning Signals. Academic Journal of Sociology and Management, 3(2), 22-32.

[12] Zhang, Y., Fan, J., & Dong, B. (2025). Deep Learning-Based Analysis of Social Media Sentiment Impact on Cryptocurrency Market Microstructure. Academic Journal of Sociology and Management, 3(2), 13-21.

[13] Zhou, Z., Xi, Y., Xing, S., & Chen, Y. (2024). Cultural Bias Mitigation in Vision-Language Models for Digital Heritage Documentation: A Comparative Analysis of Debiasing Techniques. Artificial Intelligence and Machine Learning Review, 5(3), 28-40.

[14] Zhang, Y., Zhang, H., & Feng, E. (2024). Cost-Effective Data Lifecycle Management Strategies for Big Data in Hybrid Cloud Environments. Academia Nexus Journal, 3(2).

[15] Wu, Z., Feng, E., & Zhang, Z. (2024). Temporal-Contextual Behavioral Analytics for Proactive Cloud Security Threat Detection. Academia Nexus Journal, 3(2).

[16] Ji, Z., Hu, C., Jia, X., & Chen, Y. (2024). Research on Dynamic Optimization Strategy for Cross-platform Video Transmission Quality Based on Deep Learning. Artificial Intelligence and Machine Learning Review, 5(4), 69-82.

[17] Zhang, K., Xing, S., & Chen, Y. (2024). Research on Cross-Platform Digital Advertising User Behavior Analysis Framework Based on Federated Learning. Artificial Intelligence and Machine Learning Review, 5(3), 41-54.

[18] Xiao, X., Zhang, Y., Chen, H., Ren, W., Zhang, J., & Xu, J. (2025). A Differential Privacy-Based Mechanism for Preventing Data Leakage in Large Language Model Training. Academic Journal of Sociology and Management, 3(2), 33-42.

[19] Xiao, X., Chen, H., Zhang, Y., Ren, W., Xu, J., & Zhang, J. (2025). Anomalous Payment Behavior Detection and Risk Prediction for SMEs Based on LSTM-Attention Mechanism. Academic Journal of Sociology and Management, 3(2), 43-51.

[20] Liu, Y., Feng, E., & Xing, S. (2024). Dark Pool Information Leakage Detection through Natural Language Processing of Trader Communications. Journal of Advanced Computing Systems, 4(11), 42-55.

[21] Chen, Y., Zhang, Y., & Jia, X. (2024). Efficient Visual Content Analysis for Social Media Advertising Performance Assessment. Spectrum of Research, 4(2).

[22] Wu, Z., Wang, S., Ni, C., & Wu, J. (2024). Adaptive Traffic Signal Timing Optimization Using Deep Reinforcement Learning in Urban Networks. Artificial Intelligence and Machine Learning Review, 5(4), 55-68.

[23] Chen, J., & Zhang, Y. (2024). Deep Learning-Based Automated Bug Localization and Analysis in Chip Functional Verification. Annals of Applied Sciences, 5(1).

[24] Zhang, Y., Jia, G., & Fan, J. (2024). Transformer-Based Anomaly Detection in High-Frequency Trading Data: A Time-Sensitive Feature Extraction Approach. Annals of Applied Sciences, 5(1).

[25] Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 833-842).

[26] Zolfaghari, M., Zhu, Y., Gehler, P., & Brox, T. (2021). Crossclr: Cross-modal contrastive learning for multi-modal video representations. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1450-1459).

[27] Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., ... & Wang, H. (2020). Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409.

[28] Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., & Rodrigo, R. (2022). Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9902-9912).

[29] Wu, Y., Liu, J., Gong, M., Gong, P., Fan, X., Qin, A. K., ... & Ma, W. (2023). Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. IEEE Transactions on Multimedia, 26, 1626-1638.

[30] Kim, D., Tsai, Y. H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., & Chandraker, M. (2021). Learning cross-modal contrastive features for video domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13618-13627).

[31] Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., & Wang,

S. (2023, October). Cross-modal contrastive learning for multimodal fake news detection. In Proceedings of the 31st ACM international conference on multimedia (pp. 5696-5704).

[32] Yan, L., Weng, J., & Ma, D. (2025). Enhanced TransFormer-Based Algorithm for Key-Frame Action Recognition in Basketball Shooting.

[33] Wang, Y., Wan, W., Zhang, H., Chen, C., & Jia, G. (2025). Pedestrian Trajectory Intention Prediction in Autonomous Driving Scenarios Based on Spatio-temporal Attention Mechanism.