

UniAR: A Unified Framework for Autism Recognition Enhanced by Multi-View Prompt Learning

Lei Xin*

Wuhan University
Wuhan, Hubei, China
2835838600@qq.com

Zeheng Wang*

Northeast Normal University
Changchun, Jilin, China
wangzeheng624@nenu.edu.cn

Jiayin Zhu*

The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, Guangdong, China
jzhu709@connect.hkust-gz.edu.cn

Shihong Huang

Nanjing Agricultural University
Nanjing, Jiangsu, China
23123122@stu.njau.edu.cn

Fanhu Zeng

Imperial College London
London, United Kingdom
challengezengfh@gmail.com

Changjiang Jiang

Wuhan University
Wuhan, Hubei, China
jiangcj@whu.edu.cn

Dengbo He

The Hong Kong University of Science
and Technology
Hong Kong, China
dengbohe@ust.hk

Yutao Yue

The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, Guangdong, China
yutaoyue@hkust-gz.edu.cn

Zhenglun Kong[†]

Harvard University
Boston, USA
zhenglun_kong@hms.harvard.edu

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder for which early and accurate diagnosis is critical to improving long-term developmental outcomes. However, existing ASD recognition methods are often constrained by the scarcity of diagnostic text data, forcing them to rely mainly on visual analysis and limiting their ability to model clinically meaningful semantic reasoning. To address this challenge, we propose UniAR, a unified framework enhanced by multi-granularity prompt learning for robust ASD recognition under heterogeneous data variations. Specifically, UniAR leverages a large multimodal model to generate hierarchical diagnostic descriptions at the word, phrase, and sentence levels, compensating for the lack of paired clinical reports. To align the generated semantics with visual evidence, we further design a Mixture-of-Experts-based Multi-Scale Alignment Module, which dynamically matches vector-quantized visual prototypes with semantic representations at corresponding granularities. Extensive experiments on four benchmarks covering brain MRI and facial expression scenarios show that UniAR consistently outperforms existing state-of-the-art methods, achieving average accuracies of 75.9% on MRI benchmarks and 91.6% on facial benchmarks, while improving average Accuracy on MRI benchmarks by 1.5 percentage points and average Accuracy on facial benchmarks by 1.2 percentage points over baselines. These results demonstrate that UniAR offers a robust and interpretable framework for ASD screening under semantic scarcity.

*The first three authors contributed equally to this work.

[†]Zhenglun Kong is the corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3835308>

CCS Concepts

- **Computing methodologies** → **Natural language generation**;
- **Applied computing** → **Health informatics**.

Keywords

ASD Recognition, Affective Computing, Multimodal Learning, Multi-Granularity Prompt Learning, Cross-Modal Alignment

ACM Reference Format:

Lei Xin, Zeheng Wang, Jiayin Zhu, Shihong Huang, Fanhu Zeng, Changjiang Jiang, Dengbo He, Yutao Yue, and Zhenglun Kong. 2026. UniAR: A Unified Framework for Autism Recognition Enhanced by Multi-View Prompt Learning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835308>

1 Introduction

Autism Spectrum Disorder (ASD) represents a complex neurodevelopmental condition characterized by core deficits in social communication and interaction, alongside restricted, repetitive patterns of behavior. According to statistics from the World Health Organization (WHO), the global average prevalence of ASD in children is approximately 1% [27]. A substantial body of research indicates that early childhood, typically defined as 0–6 years, serves as a critical window for neurodevelopment; achieving early identification and intervention during this phase is instrumental in improving long-term developmental outcomes [2, 28, 29]. Consequently, enhancing the accessibility and accuracy of early ASD diagnosis holds significant social and public health implications [29].

In clinical practice, screening for autism spectrum disorder (ASD) relies on standardized scales (e.g., M-CHAT-R/F) and parent reports [26]. However, these subjective measurement methods often struggle to address the inherent heterogeneity of ASD presentation [9, 21, 26, 40]. Although machine learning has been applied to automatically identify ASD through behavioral imaging, existing

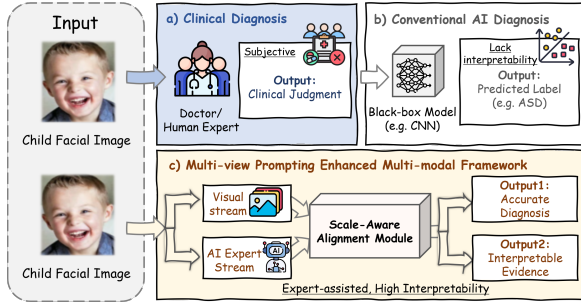


Figure 1: The framework comparisons. Existing autism recognition methods (a-b) are mainly implemented by domain experts or neural networks. In contrast, UniAR (c) integrates expert knowledge with neural networks to provide a more interpretable recognition solution, which is capable of outputting both diagnostic reports and results simultaneously.

methods primarily focus on unimodal visual analysis. Therefore, they overlook crucial semantic descriptors in clinical diagnosis, such as emotional responses and social interactions [35].

While multimodal learning [17, 44] offers a promising avenue for integrating visual and semantic data, its application faces several obstacles. First, high-quality paired visual-text data is scarce and expensive to obtain [42]. Furthermore, strict privacy regulations surrounding sensitive diagnostic text severely limit data sharing [7, 22]. These limitations prevent current multimodal methods from fully realizing their potential, thus hindering the deployment of accurate and practical diagnostic systems.

To address the challenges of data limitations and missing text modalities, we propose **UniAR, a Multi-Granularity Prompt Learning framework based on Cross-modal Alignment**, as shown in Figure 1. Specifically, we utilize generative semantic augmentation to automatically synthesize multi-granularity descriptive text from visual inputs, effectively compensating for the scarcity of real diagnostic reports. Recognizing the complexity of ASD symptoms, we design a hierarchical “word-phrase-sentence” alignment mechanism that maps semantic descriptions to pixel-level, regional, and global visual features, respectively. By linking diagnostic semantics with behavioral biomarkers at different granularities, UniAR enables effective multi-scale visual-semantic fusion.

We conduct extensive experimental validation of UniAR on four heterogeneous public benchmarks, covering brain MRI connectivity (ABIDE I & II) and facial behavioral expressions (HRM & Kaggle). While these public benchmarks mainly provide standard binary ASD-vs-control evaluation settings, they are limited in reflecting fine-grained symptom severity and real-world distribution shifts encountered in practical screening. **To address this limitation, we further construct the first cross-platform social-media autism benchmark, termed ASD-MM**, which consists of autism-related videos paired with expert-associated diagnostic descriptions and formulates autism-related assessment as a four-class problem (*i.e.*, Normal, Mild, Moderate, and Severe). Experimental results show that UniAR consistently outperforms strong single-modal and multimodal baselines, achieving average accuracies of 75.9% on MRI benchmarks and 91.6% on facial benchmarks. UniAR also

achieves the best performance on the challenging ABIDE I and dynamic HRM datasets. Moreover, on the self-constructed ASD-MM benchmark, UniAR consistently improves over the EAC baseline across all evaluated social platforms, further demonstrating its robustness and practical generalization.

In summary, our paper makes the following contributions:

- We propose UniAR, a unified visual-semantic alignment framework for robust ASD recognition that transforms low-level visual biomarkers into interpretable high-level diagnostic insights through multi-granularity prompting and semantic-guided modeling.
- We develop a hierarchical word-phrase-sentence visual-semantic alignment mechanism that associates visual biomarkers with diagnostic semantics at multiple granularities, and further enhance visual representation learning through codebook-based prototype refinement, thereby improving robustness and diagnostic reasoning under limited data.
- Extensive experiments on public MRI and facial expression benchmarks, as well as a cross-platform social-media evaluation setting, demonstrate that UniAR consistently outperforms strong baselines while showing promising robustness and semantic interpretability.

2 Related Work

ASD Recognition. Automated ASD recognition can be broadly grouped into neuroimaging-based diagnosis, facial-behavior analysis, and emerging multimodal systems. In neuroimaging, recent review papers indicate that MRI-based ASD diagnosis remains a major research paradigm, with most studies focusing on stronger discriminative modeling of resting-state or structural biomarkers on datasets such as ABIDE [7, 30, 35]. Representative methods include connectome-based predictive modeling, graph-based modeling of functional connectivity, and more recent transformer or spectral architectures for ASD-related brain analysis [1, 4, 5, 18–20, 32]. In parallel, facial and behavioral approaches identify ASD-related cues from head pose, facial expressions, and social-behavioral responses in images or videos [3, 12, 20]. Their implementations often build on robust FER techniques such as region attention, uncertainty modeling, and noisy-label learning [31, 38, 39, 46, 47]. Some work further explores multimodal evidence such as audio-visual interaction modeling [49]. Prior ASD-specific methods can be summarized as connectome modeling, behavioral cue modeling, and multimodal fusion, yet they mainly strengthen discriminative representations while explicit clinical-semantic modeling remains limited.

Semantic Reconstruction and Missing-Modality Learning.

Another related line addresses limited supervision through augmentation, modality completion, or prompt-based adaptation. Classical augmentation improves generalization by enriching the sample space [33], and recent neuroimaging work has extended this idea to fMRI-specific augmentation and computation toolchains [6]. In ASD settings, ASD-GANNet and GARL enhance robustness by synthesizing pseudo-samples or enriched feature distributions for small-data regimes [16, 50]. More broadly, missing-modality learning studies how to preserve performance when one or more modalities are absent, typically through imputation, distillation, or robust fusion [37, 42]. Prompt-based multimodal adaptation [17, 43] further

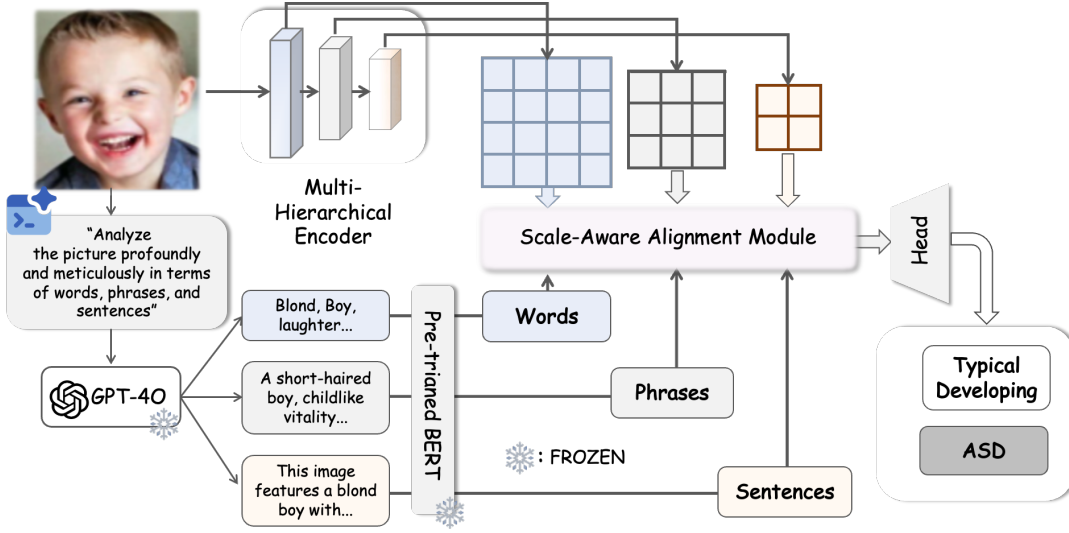


Figure 2: Overview of UniAR. The visual stream extracts multi-scale visual prototypes via a learnable visual codebook-augmented multi-hierarchical encoder. The semantic stream uses hierarchical prompts with GPT-4o to generate word-, phrase- and sentence-level descriptions, encoded by a frozen text encoder. A Scale-Aware Alignment Module aligns and fuses these multi-modal features for final classification.

shows that learnable prompts can improve cross-modal conditioning and transfer [8, 48]. DeFakerOne [34] introduces a industrial-grade image detection framework, demonstrating the feasibility leveraging massive datasets. Taken together, these methods can be summarized as sample-space augmentation, feature or modality completion, and prompt-conditioned adaptation. However, they mainly target robustness under incomplete inputs or limited labels, whereas diagnosis-level semantic reconstruction under missing clinical text remains underexplored. Fake-HR1 [14] and DS-Net [25] is based on question difficulty and the quality of dataset to achieve adaptive perception training.

Medical Visual-Language Alignment. Medical vision-language learning has progressed from coarse report-level matching to increasingly fine-grained semantic alignment. At a conceptual level, multi-grained vision-language pretraining showed the value of matching textual units with visual concepts across granularities [45]. In the medical domain, GLoRIA aligns report words with image sub-regions for label-efficient representation learning; MGCA models region-, instance-, and disease-level correspondences between medical images and reports; and MedKLIP introduces structured medical knowledge to align entity-level descriptions with image patches for diagnosis and grounding [11, 36, 41]. More recent work extends this trend to brain-domain vision-language pretraining and aspect-level disease semantic decomposition [23, 24]. Ivy-Fake [13] and TwSG [15] pioneered an explainable dataset for image understanding. Overall, existing methods can be summarized as report-level matching, multi-granularity local-global alignment, and knowledge-guided semantic grounding. However, they generally assume that paired reports or curated disease descriptions are available during training, leaving report-absent medical settings insufficiently studied.

3 The Proposed Method

3.1 Preliminaries

ASD recognition often suffers from *semantic scarcity*: visual observations are available, while fine-grained clinical descriptions are usually missing. UniAR addresses this problem by modeling visual evidence and the latent diagnostic semantics inferred from it.

Problem Formulation. We formulate ASD recognition as a supervised classification problem. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset, where $x_i \in \mathcal{X}$ is the visual input and $y_i \in \mathcal{Y}$ is the diagnostic label. Since explicit clinical semantics are typically unavailable, we approximate them using a large multimodal generator $\mathcal{G}$ and optimize $\theta^* = \arg \max_{\theta} \sum_{i=1}^N \log P(y_i | x_i, \mathcal{G}(x_i); \theta)$, where θ denotes the learnable parameters of UniAR.

Loss Function. The overall objective combines classification, semantic alignment, hierarchical regularization, fusion quality, and prototype learning: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{sep}} + \gamma \mathcal{L}_{\text{fusion}} + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}}$, where $\mathcal{L}_{\text{cls}}$ supervises ASD prediction, $\mathcal{L}_{\text{align}}$ enforces cross-modal consistency, $\mathcal{L}_{\text{sep}}$ preserves complementary hierarchical cues, $\mathcal{L}_{\text{fusion}}$ regularizes multi-scale fusion, and $\mathcal{L}_{\text{vq}}$ stabilizes prototype learning.

Based on this formulation, UniAR is designed to reconstruct missing diagnostic semantics from visual evidence and align them with multi-scale visual representations for robust ASD recognition.

3.2 Framework Overview

ASD recognition under semantic scarcity involves three coupled challenges. First, raw visual observations are highly sensitive to subject-specific appearance variations, acquisition noise, and site-specific heterogeneity, making the learned representations unstable across samples. Second, although visual data may contain discriminative behavioral biomarkers, explicit clinical semantic supervision

is typically unavailable, preventing the model from linking visual evidence to interpretable diagnostic concepts. Third, even when semantic descriptions are constructed, visual and textual representations remain mismatched in both feature distribution and abstraction level, making direct multimodal fusion suboptimal.

To address these issues, we propose UniAR, a unified visual-semantic framework for ASD recognition, as illustrated in Figure 2. The framework follows a three-stage pipeline. First, a *visual representation construction stage* transforms raw visual inputs into prototype-based multi-scale features through a hierarchical encoder and a learnable visual codebook, thereby reducing instance-specific noise while preserving ASD-relevant patterns. Second, a *semantic construction stage* leverages a large multimodal model to generate hierarchical diagnostic descriptions at the word, phrase, and sentence levels, providing explicit semantic anchors for otherwise purely visual evidence. Third, a *scale-aware alignment stage* progressively aligns visual prototypes with semantic embeddings at matched granularities, enabling robust cross-modal interaction and final ASD prediction.

In this way, UniAR does not treat multimodal learning as a simple combination of visual and textual features. Instead, it first stabilizes the visual representation, then reconstructs missing diagnostic semantics, and finally resolves the scale mismatch between the two modalities through explicit alignment. This design allows the framework to jointly improve classification robustness and diagnostic interpretability.

3.3 Multi-Granularity Visual-Semantic Representation Construction Module

Following the framework overview, we first construct the multi-granularity visual-semantic representation used by UniAR. Effective ASD recognition requires both robust visual representations and semantically grounded diagnostic cues. However, raw visual features are often entangled with subject-specific variations and irrelevant noise, while stabilized visual patterns still lack explicit clinical meaning. We therefore construct the visual-semantic representation in three stages: multi-granularity representation space, codebook-enhanced Hierarchical Visual Encoding and prompt-driven multi-granularity semantic generation.

Multi-Granularity Representation Space. To reflect the hierarchical nature of ASD evidence, both modalities are represented at multiple granularities. For the visual modality, the encoder $\mathcal{E}_{vis}$ produces:

$$\mathcal{V} = \mathcal{E}_{vis}(x) = \{\mathbf{v}^{(l)} \in \mathbb{R}^{H_l \times W_l \times D} \mid l = 1, \dots, L\}, \quad (1)$$

where $\mathbf{v}^{(l)}$ denotes the visual representation at scale l .

For the semantic modality, the generator $\mathcal{G}$ with hierarchical prompts $\mathcal{P}$ yields:

$$\mathcal{T} = \mathcal{G}(x, \mathcal{P}) = \{T_k \mid k \in \{\text{word, phrase, sent}\}\}. \quad (2)$$

Here, T_k denotes the generated diagnostic description at semantic granularity k . This formulation allows UniAR to associate visual evidence with semantic concepts at matched granularities.

Codebook-Enhanced Hierarchical Visual Encoding. Directly aligning raw visual features with language is unreliable due to

identity-specific appearance, background interference, and acquisition variation. We therefore first transform raw observations into more stable visual prototypes.

Given a visual observation, a multi-hierarchical encoder extracts feature maps at multiple scales: $\mathcal{V} = \{\mathbf{v}_l\}_{l=1}^L$, $\mathbf{v}_l \in \mathbb{R}^{H_l \times W_l \times D}$, where $\mathbf{v}_l$ denotes the feature representation at scale l . To suppress redundant variations, we introduce a learnable *Visual Codebook* $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^K$, where each codeword $\mathbf{c}_k \in \mathbb{R}^D$ represents an ASD-relevant visual prototype.

For each spatial feature vector $\mathbf{v}_l^{(i,j)}$, nearest-neighbor quantization is performed as:

$$z_q(\mathbf{v}_l^{(i,j)}) = \mathbf{c}_k, \quad \text{where } k = \arg \min_{n \in \{1, \dots, K\}} \|\mathbf{v}_l^{(i,j)} - \mathbf{c}_n\|_2. \quad (3)$$

$z_q(\cdot)$ denotes the quantization operator. By mapping continuous features into a discrete prototype space, the model suppresses instance-specific noise and encourages visually diverse samples to share ASD-salient behavioral patterns. Gradients are propagated through the quantization operation using the straight-through estimator.

Prompt-Driven Multi-Granularity Semantic Generation. Although the visual codebook stabilizes the visual stream, the resulting prototypes remain diagnostically under-specified. To provide interpretable diagnostic anchors, we further construct hierarchical semantics from the visual input.

Because real ASD datasets rarely contain dense reports, we employ GPT-4o to generate descriptions at three complementary granularities: $\mathcal{T} = \{T_{\text{word}}, T_{\text{phrase}}, T_{\text{sent}}\}$, where T_{word} , T_{phrase} , and T_{sent} denote word-level, phrase-level, and sentence-level descriptions, respectively. These levels progressively capture isolated attributes, local behavioral interactions, and holistic diagnostic context.

The generated texts are encoded by a frozen text encoder:

$$\mathbf{S}_g = \text{Enc}(T_g), \quad g \in \{\text{word, phrase, sent}\}, \quad (4)$$

yielding semantic embeddings $\mathbf{S}_{\text{word}}$, $\mathbf{S}_{\text{phrase}}$, and $\mathbf{S}_{\text{sent}}$. This hierarchical construction provides richer semantic anchors than a single global description and supports subsequent fine-grained cross-modal alignment.

3.4 Multi-Scale Alignment Module

After prototype refinement and semantic generation, the model obtains more stable visual patterns and more informative diagnostic descriptions. However, these two modalities are still not directly compatible. In particular, visual prototypes and semantic embeddings differ not only in modality, but also in abstraction level: local visual cues should not interact with sentence-level semantics in the same manner as global behavioral patterns. As a result, naive concatenation or direct fusion may introduce severe cross-modal mismatch and weaken the benefit of hierarchical semantics. To address this issue, we propose a *Multi-Scale Alignment Module (MSAM)* that explicitly performs scale-aware feature modulation and cross-modal interaction.

Cross-Modal Alignment Objective. Since visual and semantic features come from heterogeneous spaces, UniAR learns a joint representation by minimizing their latent discrepancy:

$$\min_{\theta} \sum_{l,k} D_{\phi} \left(f_v(\mathbf{v}^{(l)}), f_t(\mathbf{t}^{(k)}) \right), \quad (5)$$

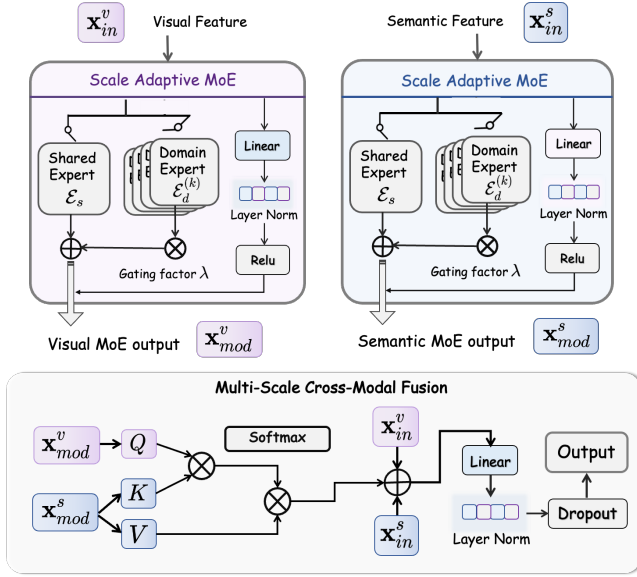


Figure 3: Detailed illustration of the Multi-Scale Alignment Module (MSAM) designed to tackle feature inconsistencies.

where $f_v(\cdot)$ and $f_s(\cdot)$ are modality-specific projection functions, and D_ϕ measures their discrepancy. This motivates the scale-aware alignment design by encouraging matched visual and semantic representations to become more compatible in the latent space.

Scale-Adaptive Modulation with MoE. Before multimodal fusion, the model should first reduce the feature discrepancy between different granularities. To this end, we introduce a scale-adaptive modulator implemented with a Mixture-of-Experts (MoE) mechanism. The main idea is to decompose each input feature into shared and scale-specific components, so that common diagnostic structure can be preserved while scale-dependent characteristics are selectively emphasized.

Let $\mathbf{x}_{in}$ denote the input feature for a specific alignment branch, which can be either a visual representation $\mathbf{x}_{in}^v$ or a semantic representation $\mathbf{x}_{in}^s$. We employ a *Shared Expert* $\mathcal{E}_s(\cdot)$ to model common patterns across different scales, and a *Scale-Specific Expert* $\mathcal{E}_d^{(k)}(\cdot)$ for the k -th granularity. The modulated feature is computed as:

$$\mathbf{x}_{mod} = \mathcal{E}_s(\mathbf{x}_{in}) + \lambda \cdot \mathcal{E}_d^{(k)}(\mathbf{x}_{in}), \quad (6)$$

where λ is a learnable gating factor. Both experts are implemented as MLPs. This design enables the model to preserve cross-scale commonality while adaptively highlighting the semantic attributes that are most relevant to each granularity, thereby producing more compatible representations for subsequent alignment.

Multi-Scale Cross-Modal Fusion. Once the features are modulated into a more compatible latent space, the next challenge is to establish effective interactions between matched visual and semantic granularities. We therefore adopt a multi-scale cross-attention mechanism to inject diagnostic semantics into visual prototypes in a scale-aware manner.

For each branch, the modulated visual features are treated as Queries, while the corresponding semantic embeddings serve as

Keys and Values. The cross-modal interaction is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{\mathbf{x}_{mod}^v (\mathbf{x}_{mod}^s)^T}{\sqrt{d_k}}\right) \mathbf{x}_{mod}^s, \quad (7)$$

where $\mathbf{x}_{mod}^v$ and $\mathbf{x}_{mod}^s$ denote the modulated visual and semantic features, respectively. Based on this interaction, the aligned representation is obtained by:

$$\mathbf{Z} = \text{LN}(\mathbf{x}_{in}^v + \mathbf{x}_{in}^s + \text{Attention}(Q, K, V)), \quad (8)$$

where $\text{LN}(\cdot)$ denotes Layer Normalization.

This process progressively injects hierarchical diagnostic semantics into the visual stream, allowing the model to connect local visual biomarkers with fine-grained linguistic cues and global visual patterns with holistic diagnostic context. Finally, the aligned outputs from all branches are aggregated to form the final representation for ASD recognition.

4 Experiments

4.1 Experimental Setting

Public Benchmarks. Autism-related biomarkers in our study are manifested in two complementary modalities, *i.e.*, brain MRI connectivity and facial behavioral expressions. To comprehensively evaluate the robustness and generalization of UniAR, we conduct experiments on four public benchmarks together with one self-constructed cross-platform social-media benchmark. *MRI Scenario (ABIDE I & II)*: We evaluate on the Autism Brain Imaging Data Exchange benchmarks. ABIDE I contains 1,112 subjects (539 ASD and 573 controls) from 17 sites, serving as a standard multi-site benchmark for ASD identification. ABIDE II includes 1,114 subjects from 19 sites and provides a more challenging setting with greater inter-site variability. *Facial Expression Scenario (HRM & Kaggle)*: We use two facial benchmarks to cover both temporal and static behavioral cues. The HRM dataset contains 131,758 frames from 1,535 videos for evaluating subtle temporal facial dynamics. In addition, we use an aggregated static facial benchmark collected from Kaggle and Zenodo, containing 12,381 images, to assess generalization.

ASD-MM Dataset Construction and Evaluation Protocol. To complement the closed binary ASD-vs-control settings of existing public benchmarks, we further construct the first cross-platform social-media autism benchmark, termed ASD-MM, for finer-grained autism assessment under realistic distribution shifts. ASD-MM consists of autism-related videos paired with expert-associated diagnostic descriptions and defines autism-related assessment as a four-class problem with labels *Normal*, *Mild*, *Moderate*, and *Severe*. The videos are collected from three mainstream social-media platforms, namely Bilibili, YouTube, and TikTok. After data cleaning and screening, representative frames are manually extracted from each video for evaluation. All samples are annotated by three clinical practitioners, and each sample is associated with an expert-written diagnostic description covering three dimensions: social competence, language expression, and behavior. Category statistics of ASD-MM are summarized in Figure 5. In experiments, ASD-MM serves as an external validation benchmark for both cross-platform four-class classification and quantitative interpretability evaluation. **Baselines.** We compare UniAR against representative methods designed for the two visual scenarios. For *MRI scenarios*, we include

Table 1: Comparative experiments on MRI scenarios.

Method	ABIDE				ABIDE II				AVG			
	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC
CPM	61.4	58.3	53.3	64.7	60.3	61.0	58.5	62.0	60.9	59.7	55.9	63.4
Braingnn	60.2	60.4	50.3	58.2	58.2	59.3	50.2	57.3	59.2	59.9	50.3	57.8
SpectBGNN	59.6	67.3	51.9	61.4	69.6	71.0	72.2	71.4	64.6	69.2	62.1	66.4
STAGIN	68.5	71.2	69.2	74.1	68.4	67.8	69.1	72.0	68.5	69.5	69.2	73.1
Bolt	67.2	68.4	66.3	71.3	70.2	70.5	71.4	75.0	68.7	69.5	68.9	73.2
Causality	69.7	54.3	73.4	74.2	71.7	65.7	74.2	75.3	70.7	60.0	73.8	74.8
BrainNetMLP	76.8	77.0	76.5	80.2	71.5	68.5	72.5	74.8	74.2	72.8	74.5	77.5
GPT-4o	77.5	78.2	77.6	81.3	71.7	68.7	72.7	74.6	74.4	73.5	75.2	78.0
Ours	78.5	79.2	78.8	82.5	73.3	69.4	73.7	76.7	75.9	74.3	76.3	79.6

Table 2: Comparative experiments on facial expression scenarios.

Method	HRM				Kaggle				AVG			
	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC
RAN	78.4	73.5	75.8	81.0	98.2	97.8	97.5	99.1	88.3	85.7	86.7	90.1
SCN	79.3	74.8	76.5	82.5	94.5	93.2	94.8	95.5	86.9	84.0	85.7	89.0
DMUE	80.5	77.1	77.5	83.9	97.1	96.5	97.8	98.2	88.8	86.8	87.7	91.1
RUL	80.8	76.9	78.4	84.2	97.6	96.8	<u>98.1</u>	98.5	89.2	86.9	<u>88.3</u>	91.4
EAC	81.2	77.5	78.9	84.8	96.3	95.9	<u>96.7</u>	97.5	88.8	86.7	<u>87.8</u>	91.2
GPT-4o	82.6	79.2	80.3	85.7	98.2	97.2	97.8	<u>98.7</u>	<u>90.4</u>	88.2	89.1	<u>92.1</u>
Ours	84.5	81.2	81.8	87.6	98.7	97.9	98.2	98.7	91.6	89.6	90.0	93.2

methods from six paradigms: (i) *graph-based learning*, including BrainGNN [19], STAGIN [18], and Spectral BGNN [4]; (ii) *temporal modeling*, represented by Bolt [1]; (iii) *classical statistical prediction*, represented by CPM [32]; (iv) *causal modeling*, represented by the causality-inspired LSTM framework [5]; and (v) *MLP-based modeling*, represented by BrainNetMLP [10], an efficient baseline for functional brain network classification. For *facial expression scenarios*, we compare with robust CNN-based, vision-language, and instruction-tuned methods, including SCN [38], DMUE [31], RUL [46], EAC [47], RAN [39]. For the cross-platform ASD-MM benchmark, we compare with EAC under the same eval protocol.

Metrics. For the public classification benchmarks, we report Accuracy, Precision, Recall, and Area Under the ROC Curve (AUC), which measure diagnostic correctness and discriminative ability. For the ASD-MM benchmark, we report Accuracy, F1-score, and Recall for four-class classification performance. To evaluate the quality of generated diagnostic descriptions, we further adopt ROUGE-1, ROUGE-2, and ROUGE-L against expert-written reports.

Implementation Details. Our framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU. For ABIDE I and II, we extract functional connectivity patterns as model input, while the facial datasets undergo standard visual preprocessing. Multi-Granularity prompts are generated using GPT-4o and then encoded by a frozen BERT encoder. For visual discretization, we use a multi-scale vector quantization module with a codebook size of 128 and an embedding dimension of 256. The quantized representations are projected back to the original feature space through a linear layer for dimensional consistency. In the MoE branch, each routing branch contains 4 experts, with hidden dimension 512 and input/output dimension 256. The gating network uses a 64-dimensional hidden layer to generate softmax routing coefficients. We train the model using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 32 for 30 epochs, with early stopping. For the public benchmarks, the train/test ratio is set to 7:3.

4.2 Comparison Results

Performance on MRI Scenarios. Table 1 reports the results on the heterogeneous ABIDE benchmarks. UniAR achieves the best overall performance on both ABIDE and ABIDE II, ranking first on all four metrics for ABIDE and obtaining the best Accuracy and AUC on ABIDE II. Compared with ASD-specific baselines, including CPM, Braingnn, SpectBGNN, STAGIN, Bolt and Causality, the advantage of UniAR is clear. In particular, relative to the causality-based model, UniAR improves Recall from 54.3% to 79.2% on ABIDE and from 60.0% to 74.3% on the average results, showing a much better balance between sensitivity and precision for screening-oriented ASD recognition. We further include BrainNetMLP and GPT-4o as stronger generic and foundation-model baselines, and UniAR still remains superior in overall Accuracy and AUC, indicating that the gain does not come merely from generated semantics, but from the joint effect of the multi-granularity visual-semantic representation construction module and the multi-scale alignment module.

Performance on Facial Expression Scenarios. Table 2 reports the results on the facial expression benchmarks. UniAR consistently achieves the best overall performance across both dynamic and static settings. Compared with ASD/facial-specific baselines, including RAN, SCN, DMUE, RUL, and EAC, the advantage of UniAR is clear on both HRM and Kaggle. In particular, on the dynamic HRM benchmark, UniAR reaches 84.5% Accuracy, outperforming EAC by 3.3 percentage points and GPT-4o by 1.9 percentage points, while also ranking first on Recall, Precision, and AUC. On the average results, UniAR improves Accuracy from 88.8% to 91.6% over EAC and from 90.4% to 91.6% over GPT-4o. These results suggest that the gain of UniAR does not come solely from generated descriptions, but from the joint effect of the multi-granularity visual-semantic representation construction module and the multi-scale alignment

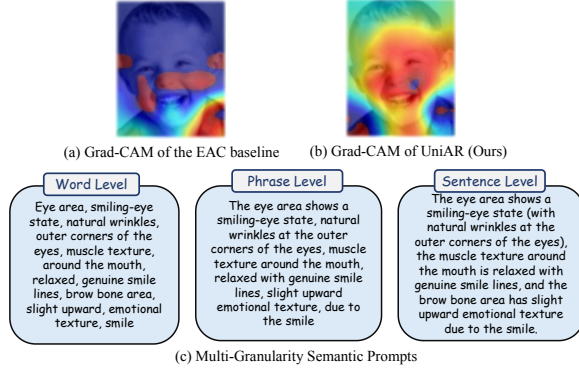


Figure 4: Qualitative visualization of attention mechanisms and hierarchical semantic alignment.

module. The consistent gains across HRM and Kaggle further indicate that the proposed framework is effective for ASD-related facial behavior understanding across both dynamic and static scenarios.

4.3 Ablation Study

We conduct ablation and design analysis from two perspectives. First, we examine the effectiveness of the proposed hierarchical alignment and fusion mechanism by progressively adding the word-level, phrase-level, sentence-level, and fusion modules. Second, we study whether two key architectural choices—the codebook size and the number of experts in the MoE branch—are reasonably selected. These aim to verify not only that the proposed components are effective, but also that the overall design of UniAR is well aligned with the demands of ASD recognition under semantic scarcity.

Module Ablation on MRI Scenarios. We first evaluate the contribution of each semantic alignment component and the fusion module on the MRI benchmarks, as shown in Table 3. Starting from the baseline (*w/o All*), performance improves steadily as word-level, phrase-level, and sentence-level alignment are introduced in sequence. This trend suggests that hierarchical semantic guidance progressively enriches the model’s understanding of ASD-related patterns, moving from fine-grained local semantics to more holistic diagnostic context. In particular, the addition of sentence-level alignment brings a clear improvement over the word+phrase setting, indicating that global semantic context is helpful for interpreting heterogeneous MRI biomarkers. After further adding the fusion module, the full model achieves the best overall performance, which supports the role of fusion in integrating multi-scale visual-semantic evidence. Similar observations can also be found on the facial expression benchmarks.

Sensitivity to Codebook Size. We next investigate the influence of codebook size on UniAR, as summarized in Table 4. Since the visual codebook is responsible for quantizing visual features into more stable prototypes, its size directly affects the balance between representation diversity and prototype compactness. The results show that a codebook size of 128 achieves the best overall performance across the four datasets. A smaller codebook (64) appears insufficient to capture diverse ASD-related visual patterns, while a

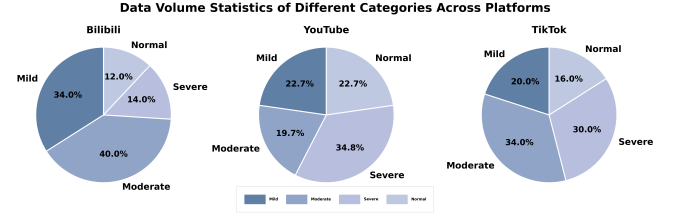


Figure 5: Data volume statistics of different categories across platforms.

larger one (256) introduces limited or inconsistent gains, suggesting that excessive prototype capacity may weaken representation compactness and stability. These results support the choice of a moderate codebook size for robust multi-scale visual abstraction.

Sensitivity to the Number of Experts. We further analyze the effect of the number of experts in the MoE-based modulation branch in Table 5, which shows that increasing the number of experts from 2 to 4 consistently improves performance across most datasets, and the best overall results are achieved with 4 experts. This suggests that a moderate increase in expert diversity is beneficial for modeling heterogeneous visual-semantic patterns, especially when aligning features from different abstraction levels. In contrast, smaller expert configurations provide limited routing flexibility and are less effective at handling the complexity of cross-dataset variations.

4.4 Qualitative Analysis and Visualization

To qualitatively examine the interpretability of UniAR, we use Grad-CAM to project decision activations back to the input space, where warmer colors indicate higher relevance. As shown in Figure 4, UniAR exhibits two visible advantages over the EAC baseline. First, its high-response regions are more concentrated on clinically relevant facial areas, such as the periorbital and perioral regions, with substantially less attention leakage to irrelevant background regions. Second, the activation patterns of UniAR appear more spatially coherent, whereas the baseline shows more scattered and fragmented responses. These observations suggest that UniAR establishes a more reliable correspondence between facial evidence and diagnostically relevant semantics.

These qualitative patterns are also consistent with the design of our hierarchical visual-semantic alignment. Word-level prompts provide fine-grained semantic anchors for local facial cues, phrase-level prompts capture relations among facial components and behavioral patterns, and sentence-level prompts introduce more global diagnostic context. Although Grad-CAM does not provide direct causal evidence for each semantic level, the improved localization and coherence of UniAR support the effectiveness of aligning visual evidence with multi-granularity diagnostic semantics.

4.5 Cross-Platform Quantitative Validation and Interpretable Analysis

To further evaluate the practical generalization of UniAR beyond closed public benchmarks, we conduct additional experiments on the self-constructed ASD-MM benchmark. In particular, this setting is designed to assess two aspects: whether UniAR remains robust

Table 3: Ablation study on MRI scenarios.

Method	Modules				ABIDE				ABIDE II				AVG			
	Word	Phrase	Sent.	Fusion	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC	Acc	Recall	Pre	AUC
Baseline	×	×	×	×	76.2	77.1	76.5	80.1	69.1	65.8	70.2	72.3	72.7	71.5	73.4	76.2
w/ Word	✓	×	×	×	76.9	77.6	77.2	80.7	70.0	66.9	71.5	73.1	73.5	72.3	74.4	76.9
w/ Word, Phrase	✓	✓	×	×	77.6	78.3	78.0	81.4	71.0	68.1	72.6	74.2	74.3	73.2	75.3	77.8
w/ Word, Phrase, Sentence	✓	✓	✓	×	78.1	78.8	78.5	82.0	71.8	68.8	73.2	75.0	75.0	73.8	75.9	78.5
Ours	✓	✓	✓	✓	78.5	79.2	78.8	82.5	73.3	69.4	73.7	76.7	75.9	74.3	76.3	79.6

Table 4: Experimental results of different codebook sizes on multiple datasets.

Codebook Size	HRM				Kaggle				ABIDE				ABIDE II			
	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC
64	83.8	80.5	81.1	86.9	98.4	97.6	97.9	98.1	78.1	78.6	78.2	81.9	71.8	68.9	73.2	75.2
128	84.5	81.2	81.8	87.6	98.7	97.9	98.2	98.7	78.5	78.9	78.8	82.5	72.3	69.4	73.7	75.7
256	84.2	81.4	81.5	87.5	98.5	97.7	98.1	98.8	78.3	78.9	78.6	82.6	72.1	69.3	73.5	75.8

Table 5: Comparative experiments with different numbers of experts on multiple datasets.

Number of Experts	HRM				Kaggle				ABIDE				ABIDE II			
	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC
2	84.2	80.9	81.5	87.2	98.4	97.7	97.9	98.4	78.2	78.8	78.6	82.2	72.2	69.2	73.5	75.3
3	84.4	81.2	81.7	87.4	98.7	97.7	98.1	98.6	78.4	78.7	78.9	82.3	72.1	69.2	73.6	75.5
4	84.5	81.2	81.8	87.6	98.7	97.9	98.2	98.7	78.5	79.2	78.8	82.5	72.3	69.4	73.7	75.7

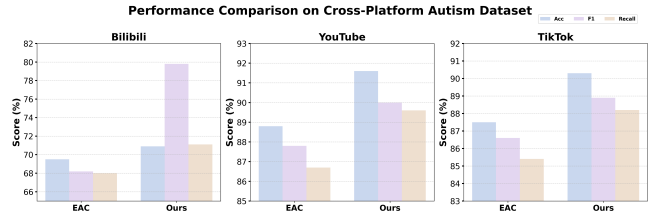
under cross-platform distribution shifts, and whether the generated diagnostic descriptions remain semantically consistent with expert-associated annotations in realistic scenarios.

Cross-Platform Four-Class Classification. We first evaluate UniAR on ASD-MM under the four-class setting of *Normal*, *Mild*, *Moderate*, and *Severe*. As shown in Figure 6, UniAR consistently outperforms the EAC baseline across the evaluated social-media platforms in terms of Accuracy, F1-score, and Recall. These results indicate that the proposed visual-semantic alignment strategy generalizes more robustly than a strong task-specific baseline under realistic cross-platform distribution shifts.

Quantitative Interpretability Evaluation. Since the diagnostic descriptions used in UniAR are generated rather than directly observed, it is important to examine whether they remain semantically faithful to expert knowledge instead of introducing unreliable hallucinated content. To this end, we further compare the generated descriptions of UniAR with expert-associated diagnostic reports using ROUGE-1, ROUGE-2, and ROUGE-L. As reported in Table 6, UniAR achieves favorable semantic consistency with expert reports across all evaluated platforms. These results suggest that the generated diagnostic descriptions capture diagnostically meaningful content to a reasonable extent, thereby providing additional quantitative support for the interpretability of UniAR in realistic deployment settings.

5 Conclusion

This paper addresses the challenge of semantic scarcity in ASD recognition by proposing UniAR, a unified framework that reconstructs hierarchical diagnostic semantics through multi-granularity

**Figure 6: Cross-platform performance comparison on social media datasets.****Table 6: Evaluation metrics on cross-platform datasets.**

Platform	Rouge-1	Rouge-2	Rouge-L
Bilibili	72.3	65.7	70.1
YouTube	78.6	72.4	76.9
TikTok	77.8	71.3	76.2

prompt learning and aligns them with visual evidence in a scale-aware manner. Extensive experiments show that structured semantic supervision is highly beneficial for robust ASD recognition, especially under heterogeneous MRI settings, subtle facial behavior variations, and incomplete diagnostic text conditions. By combining prototype-based visual representation refinement with hierarchical visual-semantic alignment, UniAR effectively improves both robustness and interpretability, providing a practical way to bridge low-level biomarkers and high-level diagnostic reasoning.

References

- [1] Hasan A Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH Dar, and Tolga Çukur. 2023. BoIT: Fused window transformers for fMRI time series analysis. *Medical image analysis* 88 (2023), 102841.
- [2] Geraldine Dawson, Emily JH Jones, Kristen Merkle, Kaitlin Venema, Rachel Lowy, Susan Faja, Dana Kamara, Michael Murias, Jessica Greenon, Jamie Winter, et al. 2012. Early behavioral intervention is associated with normalized brain activity in young children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry* 51, 11 (2012), 1150–1159.
- [3] Andong Deng, Taojiannan Yang, Chen Chen, Qian Chen, Leslie Neely, and Sakiko Oyama. 2024. Language-assisted deep learning for autistic behaviors recognition. *Smart Health* 32 (2024), 100444.
- [4] Peiyu Duan, Nicha C Dvornek, Jiyao Wang, Jeffrey Eilbott, Yuexi Du, Denis G Sukhodolsky, and James S Duncan. 2024. Spectral brain graph neural network for prediction of anxiety in children with autism spectrum disorder. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.
- [5] Peiyu Duan, Nicha C Dvornek, Jiyao Wang, Lawrence H Staib, and James S Duncan. 2025. Causal Modeling of fMRI Time-Series for Interpretable Autism Spectrum Disorder Classification. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.
- [6] Yuqi Fang, Junhao Zhang, Linmin Wang, Qianqian Wang, and Mingxia Liu. 2025. ACTION: Augmentation and computation toolbox for brain network analysis with functional MRI. *NeuroImage* 305 (2025), 120967.
- [7] Mogana Darshini Ganggayah, Diyan Zhao, Ewilly Jie Ying Liew, Nurul Aqilah Mohd Nor, Thayapari Paramasivam, Yu Ying Lee, Nurhasniza Idham Abu Hasan, and Shazwani Shaharuddin. 2025. Accelerating autism spectrum disorder care: A rapid review of data science applications in diagnosis and intervention. *Asian Journal of Psychiatry* 108 (2025), 104498.
- [8] Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13572–13586.
- [9] Hannah Hobson and Stephanie Petty. 2021. Moving forwards not backwards: heterogeneity in autism spectrum disorders. *Molecular Psychiatry* 26, 12 (2021), 7100–7101.
- [10] Jiacheng Hou, Zhenjie Song, and Ercan Engin Kuruoglu. 2025. BrainNetMLP: An Efficient and Effective Baseline for Functional Brain Network Classification. In *International Workshop on Efficient Medical Artificial Intelligence*. Springer, 11–20.
- [11] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3942–3951.
- [12] Walaa N Ismail and Mona AS Ali. 2025. Multiscale attention-based network to enhance detection and classification of autism spectrum disorders using convolutional neural network. *PeerJ Computer Science* 11 (2025), e3134.
- [13] Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. 2026. Ivy-Fake: A Unified Explainable Framework and Benchmark for Image and Video AIGC Detection. In *Proceedings of the 2026 International Conference on Multimedia Retrieval (ICMR '26)*. Association for Computing Machinery, 2438–2447. doi:10.1145/3805622.3810615
- [14] Changjiang Jiang, Xinkuan Sha, Fengchang Yu, Jingjing Liu, Jian Liu, Mingqi Fang, Chenfeng Zhang, and Wei Lu. 2026. Fake-HR1: Rethinking Reasoning of Vision Language Model for Synthetic Image Detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 10482–10486. doi:10.1109/ICASSP55912.2026.11462736
- [15] Changjiang Jiang, Qiannian Zhao, Lei Xin, Jinxiang Xie, Preslav Nakov, and Zhuohan Xie. 2026. Think with Structured Grounding: Perceptual Reinforcement Learning for Chart and Visual-Tabular Understanding. *arXiv preprint arXiv:2608.22429* (2026).
- [16] Naseer Ahmed Khan and Xuequn Shang. 2024. Asd-ganet: a generative adversarial network-inspired deep learning approach for the classification of autism brain disorder. *Brain Sciences* 14, 8 (2024), 766.
- [17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19113–19122.
- [18] Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. 2021. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems* 34 (2021), 4314–4327.
- [19] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. 2021. Brainngn: Interpretable brain graph neural network for fmri analysis. *Medical image analysis* 74 (2021), 102233.
- [20] Xuehan Liu, Md Rakibul Hasan, Tom Gedeon, and Md Zakir Hossain. 2024. MADE-for-ASD: A multi-atlas deep ensemble network for diagnosing Autism Spectrum Disorder. *Computers in Biology and Medicine* 182 (2024), 109083.
- [21] Catherine Lord. 2019. Recognising the heterogeneity of autism. *The Lancet Psychiatry* 6, 7 (2019), 551–552.
- [22] Kelly N Michelson, James G Adams, and Joshua MM Faber. 2022. Navigating clinical and business ethics while sharing patient data. *Jama* 327, 11 (2022), 1025–1026.
- [23] Masoud Monajatipoor, Zi-Yi Dou, Aichi Chien, Nanyun Peng, and Kai-Wei Chang. 2024. Medical vision-language pre-training for brain abnormalities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 11159–11164.
- [24] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. 2024. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11492–11501.
- [25] Chenfan Qu, Yiwu Zhong, Xuekang Zhu, Junchi Li, Changjiang Jiang, Lianwen Jin, et al. 2026. Detect Any AI-Counterfeited Text Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 35437–35450.
- [26] Diana L Robins, Karis Casagrande, Marianne Barton, Chi-Ming A Chen, Thyde Dumont-Mathieu, and Deborah Fein. 2014. Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics* 133, 1 (2014), 37–45.
- [27] Nader Salari, Shabnam Rasoulpoor, Shna Rasoulpoor, Shamarina Shohaimi, Sima Jafarpour, Nasrin Abdoli, Behnam Khaledi-Paveh, and Masoud Mohammadi. 2022. The global prevalence of autism spectrum disorder: a comprehensive systematic review and meta-analysis. *Italian journal of pediatrics* 48, 1 (2022), 112.
- [28] Micheal Sandbank, Kristen Bottema-Beutel, Shannon Crowley LaPoint, Jacob I Feldman, D Jonah Barrett, Nicolette Caldwell, Kacie Dunham, Jenna Crank, Suzanne Albarran, and Tiffany Woyanowski. 2023. Autism intervention meta-analysis of early childhood studies (Project AIM): updated systematic review and secondary analysis. *Bmj* 383 (2023).
- [29] Damian F Santomauro, Holly E Erskine, Ana M Mantilla Herrera, Paul Anthony Miller, Jamileh Shadid, Hailey Hagins, Isaac Yeboah Addo, Qorinah Estingtyas Sakilah Adnani, Bright Opoku Ahinkorah, Ayman Ahmed, et al. 2025. The global epidemiology and health burden of the autism spectrum: Findings from the Global Burden of Disease Study 2021. *The Lancet Psychiatry* 12, 2 (2025), 111–121.
- [30] Sijr JC Schielen, Jesper Pilmeyer, Albert P Aldenkamp, and Svitlana Zinger. 2024. The diagnosis of ASD with MRI: a systematic review and meta-analysis. *Translational psychiatry* 14, 1 (2024), 318.
- [31] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6248–6257.
- [32] Xilin Shen, Emily S Finn, Dustin Scheinost, Monica D Rosenberg, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *nature protocols* 12, 3 (2017), 506–518.
- [33] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.
- [34] Guangjian Team. 2026. Venus-DeFakerOne: Unified Fake Image Detection & Localization. *arXiv preprint arXiv:2605.14091* (2026).
- [35] Md Zassim Uddin, Md Arif Shahriar, Md Nadim Mahamood, Fady Alnajjar, Md Ilaas Pramanik, and Md Atiqur Rahman Ahad. 2024. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Engineering Applications of Artificial Intelligence* 127 (2024), 107185.
- [36] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems* 35 (2022), 33536–33549.
- [37] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Learnable cross-modal knowledge distillation for multimodal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 216–226.
- [38] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6897–6906.
- [39] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2020. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 4057–4069.
- [40] Andrea Trubanova Wiecekowsky, Lashae N Williams, Juliette Rando, Kristen Lyall, and Diana L Robins. 2023. Sensitivity and specificity of the modified checklist for autism in toddlers (original and revised) a systematic review and meta-analysis. *JAMA pediatrics* 177, 4 (2023), 373–383.
- [41] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*. 21372–21383.

- [42] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825* (2024).
- [43] Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. 2025. Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection. In *The Thirteenth International Conference on Learning Representations*.
- [44] Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Modalprompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 12137–12152.
- [45] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276* (2021).
- [46] Yuhang Zhang, Chengrui Wang, and Weihong Deng. 2021. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems* 34 (2021), 17616–17627.
- [47] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. 2022. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European conference on computer vision*. Springer, 418–434.
- [48] Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453* (2025).
- [49] Zhenhao Zhao, Eunsun Chung, Kyong-Mee Chung, and Chung Hyuk Park. 2025. AV-FOS: A Transformer-Based Audio-Visual Multi-modal Interaction Style Recognition for Children with Autism Based on the Family Observation Schedule (FOS-II). *IEEE journal of biomedical and health informatics* (2025).
- [50] Yujing Zhou, Guangbo Jia, Yingdong Ren, Yingxin Ren, Zhifeng Xiao, and Yinmei Wang. 2024. Advancing ASD identification with neuroimaging: a novel GARL methodology integrating Deep Q-Learning and generative adversarial networks. *BMC Medical Imaging* 24, 1 (2024), 186.