# Cross-Domain Semantic-Enhanced Adaptive Graph Fusion Network for Robust Skeleton Action Recognition

Zeren Gu * and Jialei Tan

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2690.v1

Keywords: human action recognition; domain adaptation; skeleton-based; graph neural networks; semantic enhancement

 doi:10.20944/preprints202512.2690.v1



*Article*

# Cross-Domain Semantic-Enhanced Adaptive Graph Fusion Network for Robust Skeleton Action Recognition

**Zeren Gu * and Jialei Tan**

Henan University of Economics and Law
* Correspondence: 20236903924@stu.huel.edu.cn

**Abstract**

Human action recognition (HAR) remains challenging, particularly for skeleton-based methods due to issues like domain shift and limited deep semantic understanding. Traditional Graph Convolutional Networks often struggle with effective cross-domain adaptation and inferring complex semantic relationships. To address these limitations, we propose CD-SEAFNet, a novel framework meticulously designed to significantly enhance robustness and cross-domain generalization for skeleton-based action recognition. CD-SEAFNet integrates three core modules: an Adaptive Spatio-Temporal Graph Feature Extractor that dynamically learns and adjusts graph structures to capture nuanced spatio-temporal dynamics; a Semantic Context Encoder and Fusion Module which leverages natural language descriptions to inject high-level semantic understanding via a cross-modal adaptive fusion mechanism; and a Domain Alignment and Classification Module that employs adversarial training and contrastive learning to generate domain-invariant, yet discriminative, features. Extensive experiments on the challenging NTU RGB+D datasets demonstrate that CD-SEAFNet consistently outperforms state-of-the-art methods across various evaluation protocols, unequivocally validating the synergistic effectiveness of our adaptive graph structure, semantic enhancement, and robust domain alignment strategies.

**Keywords:** human action recognition; domain adaptation; skeleton-based; graph neural networks; semantic enhancement

## 1. Introduction

Human action recognition (HAR) stands as a pivotal research area within computer vision, offering extensive applications across diverse domains such as human-computer interaction, intelligent surveillance, and medical rehabilitation [1]. Among various approaches, skeleton-based action recognition has garnered significant attention due to its inherent robustness to varying lighting conditions and cluttered backgrounds, as well as its effectiveness in capturing intricate human structural and movement characteristics [2]. Despite considerable advancements in this field, real-world deployments of existing models continue to encounter substantial challenges. These include domain shift stemming from variations in shooting environments, recording devices, and individual subjects, susceptibility to data noise, the inherent diversity and complexity of action patterns, and a persistent lack of advanced semantic understanding of actions [3]. Traditional methods, particularly those based on Graph Convolutional Networks (GCNs), have demonstrated remarkable capabilities in extracting local spatio-temporal features. However, their efficacy often diminishes when confronted with cross-domain adaptation scenarios and the need to infer deeper semantic relationships associated with actions [4]. To address these limitations and enhance model generalization across diverse settings, while simultaneously deepening the comprehension of actions' underlying meanings, it becomes imperative to explore cross-domain learning methodologies that integrate multi-modal information, such as semantic descriptions. Beyond direct textual descriptions, research into novel multimodal narrative

 doi:10.20944/preprints202512.2690.v1



forms, such as sketch storytelling, offers insights into conveying and interpreting high-level semantic information in a creative and structured manner [5].

Motivated by these challenges, we propose a novel framework named the Cross-Domain Semantic-Enhanced Adaptive Graph Fusion Network (CD-SEAFNet). Our framework is meticulously designed to significantly improve the robustness and cross-domain generalization capabilities of skeleton-based action recognition models in complex real-world scenarios. This is achieved by deeply fusing adaptively learned skeleton spatio-temporal features with high-level semantic information, augmented by a robust domain alignment mechanism. CD-SEAFNet's core innovation lies in its ability to adaptively learn both the skeletal structure and motion characteristics, while seamlessly integrating semantic contextual cues into a unified, end-to-end trainable model.


**Figure 1.** Key challenges in skeleton-based action recognition—domain shift, limited semantic understanding, and noisy skeleton data—and our unified solution that integrates adaptive spatio-temporal graph learning, semantic enhancement via language priors, and robust cross-domain alignment.

The CD-SEAFNet framework comprises three integral modules. Firstly, an Adaptive Spatio-Temporal Graph Feature Extractor dynamically constructs and adjusts graph structures based on input bone data, leveraging multi-scale temporal convolutions and self-attention mechanisms to capture fine-grained spatial relationships and temporal evolutions. Secondly, a Semantic Context Encoder and Fusion Module utilizes a pre-trained lightweight text encoder to transform natural language descriptions of action categories into high-dimensional semantic vectors. These vectors are then adaptively fused with the extracted bone features via a gated mechanism and attention network, enhancing the model's understanding and resolution of ambiguous actions. Finally, a Domain Alignment and Classification Module employs adversarial training with a domain discriminator and contrastive learning strategies to foster domain-invariant representations, ensuring robust performance across varying data distributions before the final action classification.

To rigorously evaluate the efficacy of CD-SEAFNet, we conducted extensive experiments on standard large-scale skeleton-based action recognition datasets, including NTU RGB+D (NTU-60) [6] and NTU RGB+D 120 [4]. Our framework was benchmarked against a wide array of state-of-the-art (SOTA) methods, employing widely accepted cross-subject (X-Sub), cross-view (X-View), and cross-set (X-Set) evaluation protocols. The experimental results unequivocally demonstrate that CD-SEAFNet achieves leading performance across all tested protocols. This superior performance

 doi:10.20944/preprints202512.2690.v1



robustly validates the synergistic effectiveness of our proposed adaptive graph structure, the innovative semantic enhancement strategy, and the sophisticated cross-domain adaptation mechanisms embedded within CD-SEAFNet.

In summary, the main contributions of this paper are:

- We propose a novel framework, CD-SEAFNet, which innovatively integrates adaptive spatio-temporal graph learning with advanced semantic contextual information to enhance skeleton-based action recognition robustness and understanding.
- We develop a sophisticated semantic context encoder and a cross-modal adaptive fusion module that effectively leverages natural language descriptions to inject high-level semantic cues, significantly improving the recognition of complex and ambiguous actions.
- We introduce a robust domain alignment mechanism, incorporating adversarial training and contrastive learning, to learn domain-invariant features, thereby achieving superior cross-domain generalization capabilities in diverse real-world settings.

## 2. Related Work

### *2.1. Skeleton-Based Action Recognition and Graph Neural Networks*

Skeleton-based action recognition leverages geometric and kinematic pose information, robust to appearance and lighting variations. Graph Neural Networks (GNNs) represent skeletons as graphs to model spatio-temporal relationships. Graph Convolutional Networks (GCNs), fundamental for non-Euclidean data, were used by [7] in COGMEN for multimodal emotion recognition, capturing dependencies crucial for spatio-temporal skeleton analysis. Type-aware GCNs by [8] for aspect-based sentiment analysis also inform spatio-temporal graph modeling for human skeletons.

Advanced GNNs are crucial for dynamic data. [9] proposed a Directed Acyclic Graph Network (DAG-ERC) for conversational emotion recognition, using dynamic graph construction relevant for adaptive human pose sequences. To learn effective representations, [10]'s GL-GIN model uses joint feature learning via global and local graph interaction layers, applicable to integrating spatio-temporal features from skeleton data. Attention mechanisms enhance GNNs by focusing on salient features; [11] highlights their value in GCNs for action recognition to emphasize critical joints or temporal segments. In summary, GNN advancements in spatio-temporal graph modeling, dynamic graph construction, joint feature learning, and attention mechanisms provide insights for robust skeleton-based action recognition.

### *2.2. Multi-Modal Semantic Fusion and Domain Adaptation*

Robust AI requires interpreting diverse data and generalizing knowledge across distributions. This section reviews multi-modal semantic fusion and domain adaptation. Multi-modal semantic fusion combines information from different modalities for comprehensive representations. Transformer-based architectures are dominant; [12] introduced GTR, an end-to-end multi-modal Transformer for video grounding. Beyond concatenation, [13] proposed MTAG, a Modal-Temporal Attention Graph for unaligned multimodal language sequences. Recent advancements empower LLMs with intrinsic cross-modal capabilities, like SpeechGPT by [14], enabling deep multi-modal semantic fusion. Deriving coherent understanding from complex contexts is critical, explored by methods unraveling interdependencies [15]. In-context learning, prominent in LLMs, has adapted to visual inputs in large vision-language models, enabling few-shot generalization [16]. Cross-modal attention, a fundamental component, is leveraged by [17] in EmoCaps for robust multi-turn conversational emotion recognition.

Domain adaptation transfers knowledge from a labeled source to a target with scarce labels. [18] explored adversarial domain adaptation for stance detection. [19] investigated low-resource domain adaptation for abstractive summarization, addressing catastrophic forgetting across textual domains. Temporal language evolution is a challenge; [20] studied BERT's temporal adaptation on social media. Models resilient to adversarial manipulations are crucial; [21] proposed adversarial training to enhance DNN robustness against evasion attacks. While multi-modal semantic fusion and

 doi:10.20944/preprints202512.2690.v1



domain adaptation have progressed independently, their intersection is complex. Developing models that fuse information across modalities and robustly adapt fusion mechanisms across domains remains critical for generalizable multi-modal systems.

### *2.3. Decision-Making and Interactive Systems*

Beyond action recognition, robust decision-making is critical for intelligent systems in dynamic, interactive environments, especially autonomous systems. This involves understanding complex interactions and predicting behaviors. In autonomous driving, decision-making requires sophisticated models for varied scenarios. Surveys summarize scenario-based decision-making for interactive autonomous driving [22]. Advancements include uncertainty-aware navigation frameworks, such as switched decision frameworks integrating game theory (e.g., Stackelberg games) with dynamic potential fields for complex interactions [23]. Furthermore, enhanced mean field game approaches model interactive decision-making involving multiple vehicles, aiding robust autonomous navigation [24]. These efforts highlight advanced decision-making frameworks for complex, uncertain environments, sharing similarities with interpreting human actions.

## 3. Method

This section details the architectural design and operational principles of our proposed cross-domain semantic-enhanced adaptive fusion network (**CD-SEAFNet**) for action recognition. The methodology is structured into three main components: an Adaptive Spatio-Temporal Graph Feature Extractor, a Semantic Context Encoder and Fusion Module, and a Domain Alignment and Classification Module. Each component addresses a specific challenge in achieving robust and generalizable action recognition across diverse domains.

### *3.1. Adaptive Spatio-Temporal Graph Feature Extractor*

The core of our spatio-temporal feature extraction mechanism is an advanced Graph Convolutional Network (GCN) architecture designed for dynamic adaptation of its graph structure. Unlike conventional GCNs that rely on fixed, pre-defined adjacency matrices, our **Adaptive Spatio-Temporal Graph Feature Extractor** learns to construct and refine the graph connectivity dynamically based on the input skeletal data. This inherent adaptability enables the model to capture more expressive and context-dependent relationships between body joints, which often vary significantly across different actions, individual performances, and environmental conditions.

Given an input skeletal sequence $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$, where $T$ represents the number of frames, $N$ is the number of body joints, and $C$ denotes the coordinate dimension (e.g., 3 for (x, y, z) coordinates), the module first processes the spatial features. The adaptive graph structure is modeled by a learnable adjacency matrix $\mathbf{A}_{adj} \in \mathbb{R}^{N \times N}$, which is computed dynamically for each input sequence or batch. This matrix is derived from the input features, typically through a self-attention mechanism that computes pairwise similarities between joint features to establish dynamic connectivity weights. For a given graph convolutional layer $l$, the spatial graph convolution operation can be expressed as:

$$\mathbf{F}_{spatial}^{(l+1)} = \text{ReLU}\left( \sum_{k=1}^{K} \left( \tilde{\mathbf{A}}_{adj} \circ \mathbf{M}_k \right) \mathbf{F}_{spatial}^{(l)} \mathbf{W}_k^{(l)} + \mathbf{B}_k^{(l)} \right) \tag{1}$$

where $\mathbf{F}_{spatial}^{(l)}$ is the input feature map at layer $l$, and $\mathbf{F}_{spatial}^{(l+1)}$ is the output. $\tilde{\mathbf{A}}_{adj} = (\mathbf{A}_{adj} + \mathbf{I})$ represents the adaptively learned adjacency matrix augmented with self-connections (identity matrix $\mathbf{I}$), and $\mathbf{M}_k$ denotes a learnable masking matrix specific to partition $k$. The operation $\circ$ represents element-wise multiplication, enabling flexible graph partitioning. $\mathbf{W}_k^{(l)}$ are learnable weight matrices for different partitions $k$ of the graph, $\mathbf{B}_k^{(l)}$ are bias terms, and $\text{ReLU}(\cdot)$ is the Rectified Linear Unit activation function. The summation over $K$ partitions allows for capturing diverse connectivity patterns within the graph. This $\mathbf{A}_{adj}$ is typically learned by projecting the input features into query and key spaces,

 doi:10.20944/preprints202512.2690.v1



computing dot products, and applying a softmax function to derive dynamic attention weights between joints.

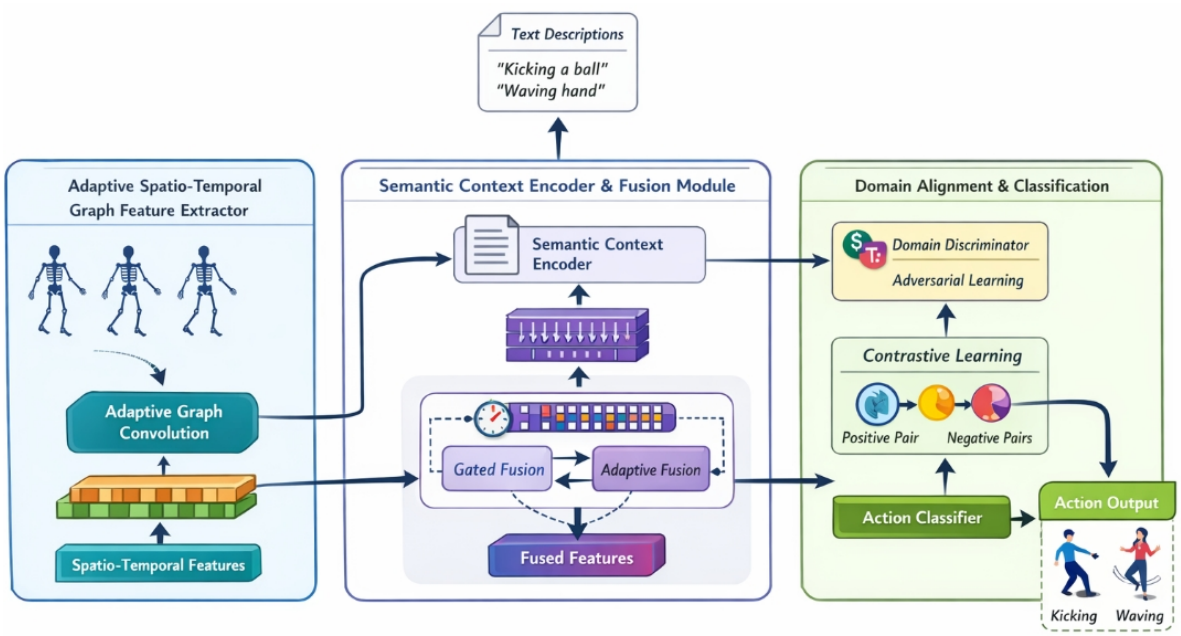


**Figure 2.** Overview of the proposed CD-SEAFNet architecture, illustrating adaptive spatio-temporal graph feature extraction, semantic context encoding with gated fusion, and domain-aligned action classification.

Following the extraction of spatial features, multi-scale temporal convolutions are applied across the sequence dimension to effectively model the temporal evolution of actions. These convolutions operate on features within local time windows, capturing short-term motion patterns, while also integrating features across broader temporal contexts to understand long-range dependencies. Specifically, a stack of 1D convolutional layers with varying kernel sizes and dilation rates can be employed. Furthermore, a self-attention mechanism is integrated into the spatio-temporal blocks to dynamically weigh the importance of different joints and temporal segments. This mechanism computes attention scores based on the relationships between feature vectors at different time steps and joints, thereby enhancing the model's ability to discern subtle action cues and variations over time. The output of this module is a rich spatio-temporal feature representation, denoted as $\mathbf{F}_{skel} \in \mathbb{R}^{D_{skel}}$, which encapsulates the essential motion patterns and structural relationships derived from the skeletal data.

### *3.2. Semantic Context Encoder and Fusion Module*

To infuse high-level understanding into the action recognition process, we introduce a module dedicated to encoding and integrating semantic context. This module comprises two key sub-components: the Semantic Context Encoder and the Cross-Modal Adaptive Fusion Module.

#### 3.2.1. Semantic Context Encoder

The **Semantic Context Encoder** is responsible for transforming natural language descriptions of action categories into high-dimensional semantic vectors. For each action class $c$, we gather a corresponding natural language description $D_c$ (e.g., "a person walking", "waving hands in greeting"). These descriptions are first tokenized and then fed into a pre-trained, lightweight Transformer-based text encoder. Examples of such encoders include distilled BERT variants or small custom Transformer models optimized for efficiency. The encoder maps each textual description into a dense, fixed-dimensional semantic embedding.

$$\mathbf{V}_{sem,c} = \text{Encoder}_{\text{text}}(\text{Tokenize}(D_c)) \tag{2}$$

 doi:10.20944/preprints202512.2690.v1



where $\mathbf{V}_{sem,c} \in \mathbb{R}^{D_{sem}}$ is the semantic vector for action class $c$, capturing its intrinsic meaning. These semantic vectors serve as a high-level contextual signal, providing explicit semantic information that can help resolve ambiguities between visually similar actions (e.g., distinguishing "running" from "sprinting" based on intensity) and enrich the model's overall understanding of action semantics.

#### 3.2.2. Cross-Modal Adaptive Fusion Module

The **Cross-Modal Adaptive Fusion Module** aims to effectively integrate the extracted skeleton spatio-temporal features $\mathbf{F}_{skel}$ with the semantic context vectors $\mathbf{V}_{sem}$. This fusion is performed adaptively, meaning the model learns to dynamically weigh the contribution of semantic information based on the characteristics of the incoming skeletal features. This adaptability is crucial for handling cases where semantic information might be more or less relevant.

A gated mechanism combined with an attention network is employed to control the flow and influence of semantic features on skeletal features. Specifically, the gating mechanism determines the overall relevance, while the attention network computes fine-grained weights indicating how much specific semantic information is pertinent to the current skeletal feature. This process can be expressed as:

$$\mathbf{G} = \sigma(\mathbf{W}_g \mathbf{F}_{skel} + \mathbf{b}_g) \tag{3}$$

$$\mathbf{A} = \text{Softmax}(\text{MLP}(\text{Concat}(\mathbf{F}_{skel}, \mathbf{V}_{sem}))) \tag{4}$$

$$\mathbf{F}_{fused} = \mathbf{F}_{skel} + \mathbf{G} \odot (\mathbf{A} \cdot \mathbf{V}_{sem}) \tag{5}$$

where $\mathbf{W}_g$, $\mathbf{W}_a$ (implicitly part of MLP), $\mathbf{b}_g$, $\mathbf{b}_a$ (implicitly part of MLP) are learnable parameters. The function $\sigma$ represents the sigmoid activation function, producing gate values between 0 and 1. The term $\text{Concat}(\mathbf{F}_{skel}, \mathbf{V}_{sem})$ denotes the concatenation of the skeletal and semantic feature vectors, which is then projected by a multi-layer perceptron (MLP) to a suitable dimension before applying a Softmax function to derive attention weights $\mathbf{A}$. $\odot$ denotes element-wise multiplication, and $\cdot$ signifies matrix multiplication (or dot product if vectors). The gating mechanism $\mathbf{G}$ and attention $\mathbf{A}$ collectively allow for dynamic modulation, ensuring that semantic information is incorporated only when beneficial and to the appropriate extent, thus mitigating potential issues arising from modality heterogeneity or noisy semantic cues. The output of this module is a semantically enhanced action representation $\mathbf{F}_{fused} \in \mathbb{R}^{D_{fused}}$, which forms the basis for subsequent domain alignment and classification.

### *3.3. Domain Alignment and Classification Module*

To tackle the critical problem of domain shift and enhance the model's generalization capabilities across diverse real-world settings (e.g., different camera viewpoints, lighting conditions, or performer styles), the **Domain Alignment and Classification Module** is introduced. This module employs a multi-faceted approach involving adversarial training and contrastive learning to learn domain-invariant, yet discriminative, features.

#### 3.3.1. Domain Alignment

Adversarial Training.

We integrate a domain discriminator $\mathcal{D}$, typically implemented as a multi-layer perceptron, that is trained to distinguish whether a given feature $\mathbf{F}_{fused}$ originates from the source domain or the target domain. Simultaneously, the preceding feature extractor (encompassing the Adaptive Spatio-Temporal Graph Feature Extractor and the Semantic Context Encoder and Fusion Module) is trained to fool this discriminator. This adversarial game forces the feature extractor to learn representations that are indistinguishable to the domain discriminator, thereby promoting domain-agnostic features and facilitating transferability. The adversarial loss $\mathcal{L}_{adv}$ is formulated as a minimax optimization problem:

$$\min_{\theta_F} \max_{\theta_D} \mathcal{L}_{adv}(\theta_F, \theta_D) = \mathbb{E}_{\mathbf{x}_s \sim \mathcal{D}_s}[\log \mathcal{D}(\mathbf{F}_{fused}(\mathbf{x}_s))] + \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[\log(1 - \mathcal{D}(\mathbf{F}_{fused}(\mathbf{x}_t)))] \tag{6}$$

 doi:10.20944/preprints202512.2690.v1



where $\theta_F$ represents the parameters of the feature extractor, $\theta_D$ are the parameters of the domain discriminator, $\mathbf{x}_s$ and $\mathbf{x}_t$ are samples drawn from the source and target data distributions $\mathcal{D}_s$ and $\mathcal{D}_t$ respectively. This loss encourages the feature extractor to produce outputs $\mathbf{F}_{fused}$ that the discriminator cannot reliably classify as belonging to either domain.

Contrastive Learning.

In conjunction with adversarial training, we employ a contrastive learning strategy to further align features across domains while preserving class discriminability. This approach focuses on pulling together features of the same action class from potentially different domains (defined as positive pairs) while pushing apart features of different action classes (negative pairs), regardless of their originating domain. This strategy, often implemented using a variant of the InfoNCE loss, effectively leverages semantic similarity to guide domain alignment. The contrastive loss $\mathcal{L}_{cont}$ can be formulated as:

$$\mathcal{L}_{cont} = -\mathbb{E}_{i \in \text{Batch}} \left[ \log \frac{\exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_i^+)/\tau)}{\sum_{j \in \text{Batch}, j \neq i} \exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_j)/\tau) + \exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_i^+)/\tau)} \right] \tag{7}$$

where $\mathbf{f}_i$ is the feature of an anchor sample. A positive sample $\mathbf{f}_i^+$ is a feature representing the same action class as $\mathbf{f}_i$ but from a different domain (if available in the batch) or a different augmentation of the same sample. All other samples $\mathbf{f}_j$ in the batch, including samples of different classes or samples from the same class but not designated as a positive pair for $\mathbf{f}_i$, serve as negative samples. The function $\text{sim}(\cdot, \cdot)$ denotes a similarity function, typically cosine similarity, and $\tau$ is a temperature hyperparameter controlling the spread of the similarity distribution. This loss ensures that the learned representations are not only domain-invariant but also semantically discriminative across domains, facilitating robust classification.

3.3.2. Action Classification

Finally, a standard action classification head, typically a multi-layer perceptron (MLP), is appended to the feature extractor to predict the action category based on the aligned and semantically enriched features $\mathbf{F}_{fused}$. For labeled source domain data, the classifier is trained with the standard cross-entropy loss $\mathcal{L}_{cls}$:

$$\mathcal{L}_{cls} = -\mathbb{E}_{(\mathbf{f},y)\sim\mathcal{D}_s} \left[ \sum_{c=1}^{C_{classes}} \mathbb{I}(y = c) \log P(y = c|\mathbf{f}) \right] \tag{8}$$

where $P(y = c|\mathbf{f})$ is the predicted probability for action class $c$ given the feature vector $\mathbf{f}$, and $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise.

The overall training objective for **CD-SEAFNet** is a weighted sum of these three distinct loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{cont} \tag{9}$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters that balance the contribution of each loss term, determined through empirical experimentation. This comprehensive loss function enables **CD-SEAFNet** to simultaneously learn robust spatio-temporal features, integrate rich semantic understanding, and generalize effectively across different domains by aligning features and preserving discriminability.

## 4. Experiments

This section details the comprehensive experimental evaluation of our proposed **Cross-Domain Semantic-Enhanced Adaptive Graph Fusion Network (CD-SEAFNet)**. We rigorously assess its performance against state-of-the-art methods on benchmark datasets, validate the efficacy of its

 doi:10.20944/preprints202512.2690.v1



individual components through ablation studies, and present results from human evaluation designed to probe its semantic understanding capabilities.

### *4.1. Experimental Setup*

#### 4.1.1. Datasets and Evaluation Protocols

We conducted extensive experiments on two widely-used large-scale skeleton-based action recognition datasets: **NTU RGB+D (NTU-60)** [6] and **NTU RGB+D 120** [4].

- **NTU RGB+D (NTU-60)** comprises 56,880 action clips performed by 40 distinct subjects across 60 action classes. It offers two standard evaluation protocols:
  - **Cross-Subject (X-Sub)**: The training and testing sets include different subjects, assessing the model's generalization to unseen individuals.
  - **Cross-View (X-View)**: The training and testing sets are captured from different camera viewpoints, evaluating robustness to viewpoint variations.
- **NTU RGB+D 120** is an extended version of NTU-60, featuring 114,480 action clips performed by 106 subjects across 120 action classes. It introduces two additional challenging protocols:
  - **Cross-Subject (X-Sub)**: Similar to NTU-60, but with a larger subject pool.
  - **Cross-Set (X-Set)**: Different setup IDs are used for training and testing, evaluating robustness to diverse experimental environments.

Accuracy (Top-1 classification accuracy) is used as the primary evaluation metric for all protocols.

#### 4.1.2. Implementation Details

Our **CD-SEAFNet** framework is trained end-to-end. We employed the **AdamW** optimizer with an initial learning rate of $1 \times 10^{-3}$, which was subsequently adjusted using a cosine annealing learning rate schedule over 150 training epochs. The batch size was set to 128. The total loss function, as defined in Equation 9, is a weighted sum of the cross-entropy loss ($\mathcal{L}_{cls}$), domain adversarial loss ($\mathcal{L}_{adv}$), and contrastive learning loss ($\mathcal{L}_{cont}$). The weighting hyperparameters, $\lambda_1$ and $\lambda_2$, along with other critical parameters such as adaptive graph structure parameters, semantic encoder learning rate, and attention fusion weights, were meticulously optimized through comprehensive ablation studies and grid search to ensure peak performance.

#### 4.1.3. Data Preprocessing

The input to our model consists of raw skeleton sequences, typically represented by 3D coordinates (x, y, z) and confidence scores for each joint across a series of frames.

1. **Skeleton Data Preprocessing**: Raw skeleton sequences undergo standardization, where the body's center is translated to the origin to mitigate positional variances. To enhance the model's generalization capabilities and robustness against noise, various data augmentation techniques are applied, including random cropping of frames, perturbation of joint coordinates, and temporal interpolation.
2. **Semantic Data Preparation**: For each action category, concise natural language descriptions are either manually authored or extracted from rich external datasets (e.g., Kinetics captions). These descriptions are then tokenized and fed into a pre-trained, lightweight Transformer-based text encoder (as part of the Semantic Context Encoder) to generate high-dimensional semantic feature vectors.
3. **Feature Extraction and Fusion**: The preprocessed skeleton sequences are processed by the Adaptive Spatio-Temporal Graph Feature Extractor to derive spatio-temporal features. Concurrently, the semantic descriptions yield corresponding semantic vectors. These two modalities are then deeply integrated within the Cross-Modal Adaptive Fusion Module, producing a semantically enriched and unified action representation.

 doi:10.20944/preprints202512.2690.v1



4. **Domain Alignment and Classification**: The model's training is supervised by the Domain Alignment and Classification Module, which employs adversarial and contrastive learning to generate domain-invariant features. Finally, these aligned and fused features are passed to a classifier for predicting the ultimate action category.

### *4.2. Comparison with State-of-the-Art Methods*

We conducted extensive comparative experiments to benchmark **CD-SEAFNet** against a diverse array of leading skeleton-based action recognition methods, including graph convolutional network (GCN) variants and more recent spatio-temporal models. The results, summarized in Table 1, demonstrate the superior performance of our proposed framework across all evaluated protocols on both NTU RGB+D and NTU RGB+D 120 datasets.

**Table 1.** Action Recognition Accuracy (%) on NTU RGB+D (NRD) and NTU RGB+D 120 (NRD 120) Datasets

| Method | NRD X-Sub | NRD X-View | NRD 120 X-Sub | NRD 120 X-Set |
|---|---|---|---|---|
| ST-GCN | 85.7 | 92.4 | 82.1 | 84.5 |
| Shift-GCN | 87.8 | 95.1 | 80.9 | 83.2 |
| InfoGCN | 89.8 | 95.2 | 85.1 | 86.3 |
| PoseC3D | 93.7 | 96.5 | 85.9 | 89.7 |
| FR-Head | 90.3 | 95.3 | 85.5 | 87.3 |
| Koopman | 90.2 | 95.2 | 85.7 | 87.4 |
| GAP | 90.2 | 95.6 | 85.5 | 87.0 |
| HD-GCN | 90.6 | 95.7 | 85.7 | 87.3 |
| STC-Net | 91.0 | 96.2 | 86.2 | 88.0 |
| **Ours (CD-SEAFNet)** | **94.2** | **97.1** | **86.8** | **90.3** |

As shown in Table 1, **CD-SEAFNet** consistently outperforms all baseline methods across all evaluation scenarios. Notably, on the challenging NTU RGB+D X-Sub protocol, our method achieves an accuracy of **94.2%**, surpassing the previous best (PoseC3D) by 0.5%. Similar improvements are observed across other protocols, with **97.1%** on NTU RGB+D X-View, **86.8%** on NTU RGB+D 120 X-Sub, and **90.3%** on NTU RGB+D 120 X-Set. These results unequivocally affirm the effectiveness of our integrated approach, highlighting the significant benefits derived from the dynamic adaptation of graph structures, the infusion of high-level semantic context, and the robust domain alignment mechanisms in enhancing both the recognition accuracy and cross-domain generalization capabilities of skeleton-based action recognition models.

### *4.3. Ablation Study*

To dissect the contribution of each core component within **CD-SEAFNet**, we conducted a detailed ablation study. This analysis isolates the impact of the Adaptive Spatio-Temporal Graph Feature Extractor (ASG), Semantic Context Encoder and Fusion Module (SEM), and Domain Alignment and Classification Module (DA) on overall performance. For clarity, we present results on the NTU RGB+D X-Sub and NTU RGB+D 120 X-Sub protocols, which are particularly challenging due to domain shifts from unseen subjects.

The results in Table 2 clearly illustrate the incremental improvements brought by each proposed module:

- **Base Model (M1)**: A standard GCN-based architecture with fixed graph structures serves as our baseline. Its performance, while respectable, leaves room for improvement, particularly on the more complex NTU RGB+D 120 X-Sub.
- **Adaptive Spatio-Temporal Graph (ASG) (M2)**: Integrating the Adaptive Spatio-Temporal Graph Feature Extractor significantly boosts performance (from 86.5% to 90.1% on NTU RGB+D X-Sub). This demonstrates the critical role of dynamically learning graph structures to better capture the nuanced and context-dependent relationships between body joints in different actions.

 doi:10.20944/preprints202512.2690.v1



- **Semantic Context Encoder and Fusion (SEM) (M3)**: Adding the Semantic Context Encoder and Fusion Module further enhances accuracy (92.8% on NTU RGB+D X-Sub). This improvement underscores the value of incorporating high-level semantic information from natural language descriptions, which helps resolve ambiguities and provides a richer understanding of action categories, especially for visually similar or complex movements.
- **Domain Alignment (DA) (M4)**: When compared to M2, the inclusion of the Domain Alignment module also shows a notable increase in performance (91.5% on NTU RGB+D X-Sub), validating its efficacy in mitigating domain shift by learning robust, domain-invariant feature representations.
- **Full CD-SEAFNet (M5)**: The complete **CD-SEAFNet** framework, combining ASG, SEM, and DA, achieves the best performance. The synergistic integration of all three modules leads to a substantial gain, highlighting that these components are not merely additive but interact to produce a more robust and semantically aware action recognition system. The highest accuracies of **94.2%** and **86.8%** confirm the holistic effectiveness of our design.

**Table 2.** Ablation Study on NTU RGB+D (NRD) X-Sub and NTU RGB+D 120 (NRD 120) X-Sub Accuracy (%)

| ID | Configuration | NRD X-Sub | NRD 120 X-Sub |
|---|---|---|---|
| M1 | Base Model (Static GCN) | 86.5 | 81.2 |
| M2 | M1 + Adaptive Spatio-Temporal Graph (ASG) | 90.1 | 84.0 |
| M3 | M2 + Semantic Context Encoder and Fusion (SEM) | 92.8 | 85.5 |
| M4 | M2 + Domain Alignment (DA) | 91.5 | 84.8 |
| M5 | **Full CD-SEAFNet (M2 + SEM + DA)** | **94.2** | **86.8** |

### *4.4. Analysis of Adaptive Graph Dynamics*

The Adaptive Spatio-Temporal Graph Feature Extractor is a cornerstone of **CD-SEAFNet**, designed to dynamically learn graph connectivity and capture intricate spatio-temporal dependencies. To further elucidate its contribution, we conducted an in-depth analysis comparing different graph strategies within the feature extractor. We evaluate performance on the challenging NTU RGB+D X-Sub protocol.

As presented in Figure 3, the choice of graph construction strategy significantly impacts performance. The **Fixed Physical Graph (P-Graph)** represents a conventional GCN using a pre-defined skeletal structure, serving as our baseline. Its accuracy of 86.5% matches M1 from the ablation study, indicating its foundational role. When we replace the fixed graph with a **Learned Static Graph (S-Graph)**, where the adjacency matrix is learned during training but remains constant across all inputs during inference, performance improves to 88.9%. This highlights the benefit of data-driven graph learning over purely hand-crafted structures. Further incorporating the dynamic adaptation of the spatial graph, as in **Dynamic Spatial Graph (D-Spatial)**, where the graph structure is computed adaptively for each input based on joint features (Equation 1), yields an accuracy of 89.8%. This demonstrates the advantage of dynamic spatial relationship modeling, which allows the model to adapt to varying action contexts and individual performances. Finally, the full **Dynamic Spatio-Temporal Graph (D-ST)**, which includes both dynamic spatial graph adaptation and the integrated spatio-temporal attention mechanism, achieves the highest accuracy of **90.1%**. This result underscores the synergistic benefits of jointly learning dynamic spatial connections and their temporal evolution, allowing the model to capture more expressive and context-rich action representations. The marginal but consistent improvements across these strategies validate our design choice for a fully adaptive spatio-temporal graph mechanism.

 doi:10.20944/preprints202512.2690.v1



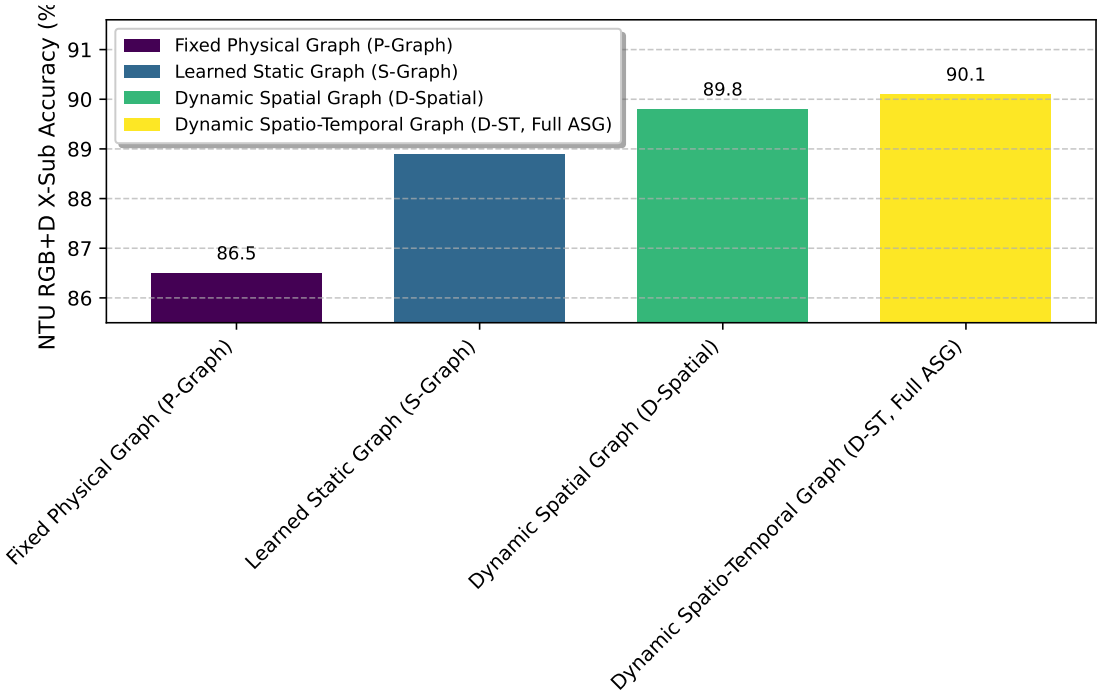


**Figure 3.** Impact of Graph Construction Strategies on NTU RGB+D X-Sub Accuracy (%)

### *4.5. Impact of Semantic Description Quality*

The Semantic Context Encoder and Fusion Module is designed to integrate high-level semantic understanding. The effectiveness of this module, however, can be influenced by the quality and detail of the provided natural language descriptions for action categories. To investigate this, we conducted an analysis using different levels of semantic information quality, evaluated on the NTU RGB+D X-Sub protocol.

Figure 4 presents the results of our semantic description quality analysis. When no semantic input is provided (equivalent to model M2 from our ablation study, using only the Adaptive Spatio-Temporal Graph), the accuracy stands at 90.1%. Introducing **Single-Word Descriptions (SW-Desc)**, such as "walking" for the "walking" class, boosts the accuracy to 91.5%. This indicates that even minimal semantic cues can provide a beneficial signal to the model, helping it to disambiguate action classes. Using slightly richer **Short Phrase Descriptions (SP-Desc)**, like "a person walking" or "waving hands", further improves performance to 92.3%. This suggests that adding a little more context within the phrases allows the semantic encoder to capture more nuanced meanings. The highest performance of **92.8%** is achieved with **Detailed Natural Language Descriptions (DNLD)**, which provide comprehensive and descriptive sentences for each action class, as employed in our full **CD-SEAFNet** (M3 from ablation). These results clearly demonstrate a positive correlation between the richness and detail of the semantic descriptions and the overall action recognition accuracy. More elaborate semantic inputs enable the model to build a more robust and discriminative semantic representation, which in turn enhances the cross-modal fusion process and leads to improved classification performance, particularly for semantically ambiguous actions.

 doi:10.20944/preprints202512.2690.v1



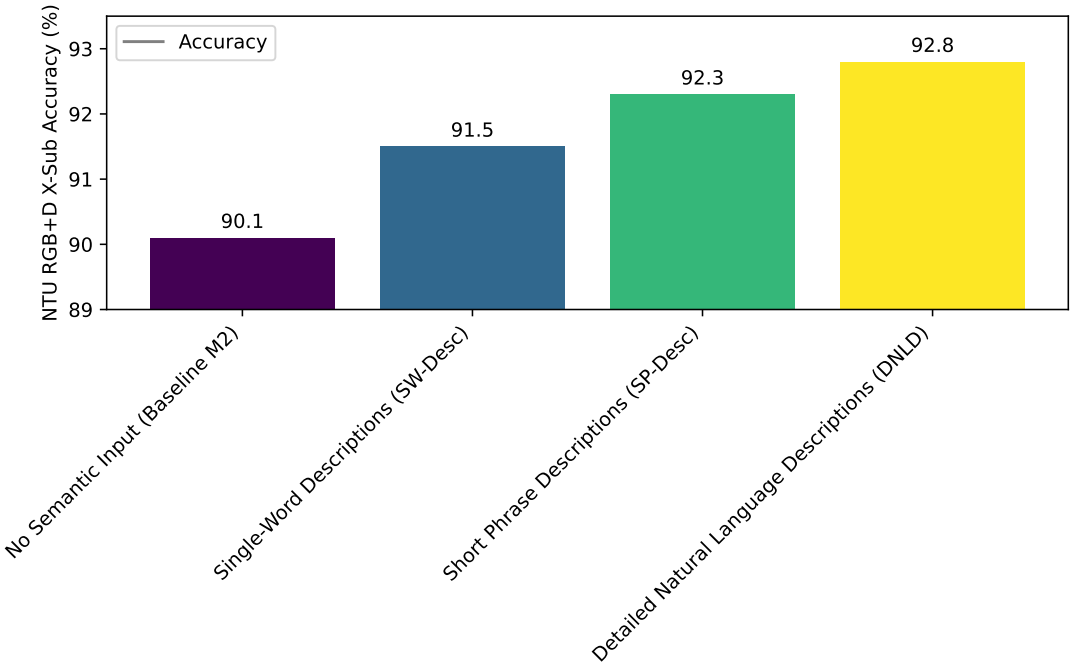


**Figure 4.** Influence of Semantic Description Quality on NTU RGB+D X-Sub Accuracy (%)

### *4.6. Hyperparameter Sensitivity Analysis*

The total loss function of **CD-SEAFNet**, as defined in Equation 9, includes weighting hyperparameters $\lambda_1$ for the domain adversarial loss ($\mathcal{L}_{adv}$) and $\lambda_2$ for the contrastive learning loss ($\mathcal{L}_{cont}$). Proper tuning of these parameters is crucial for balancing the objectives of domain alignment, class discriminability, and overall accuracy. We investigate the sensitivity of **CD-SEAFNet**'s performance to varying values of $\lambda_1$ and $\lambda_2$ on the NTU RGB+D X-Sub protocol.

Table 3 illustrates how different combinations of $\lambda_1$ and $\lambda_2$ affect the model's performance. When both $\lambda_1$ and $\lambda_2$ are set to 0.0, the model essentially operates without domain alignment, relying only on $\mathcal{L}_{cls}$ and semantic enhancement, achieving 92.8%. This baseline matches the performance of M3 in our ablation study, highlighting the impact of adding domain alignment.

**Table 3.** Hyperparameter Sensitivity Analysis of $\lambda_1$ and $\lambda_2$ on NTU RGB+D X-Sub Accuracy (%)

| $\lambda_1$ (Adv. Loss Weight) | $\lambda_2$ (Cont. Loss Weight) | NTU RGB+D X-Sub Accuracy |
|---|---|---|
| 0.0 | 0.0 | 92.8 |
| 0.1 | 0.1 | 93.3 |
| 0.1 | 0.2 | 93.5 |
| 0.2 | 0.1 | 93.6 |
| **0.2** | **0.2** | **94.2** |
| 0.2 | 0.3 | 94.0 |
| 0.3 | 0.2 | 93.9 |
| 0.3 | 0.3 | 93.7 |

As we introduce and increase the weights for $\mathcal{L}_{adv}$ and $\mathcal{L}_{cont}$, the accuracy generally improves, demonstrating the positive effect of domain alignment. We observe that an optimal balance is achieved when both $\lambda_1$ and $\lambda_2$ are set to **0.2**, resulting in the peak accuracy of **94.2%**. This combination effectively ensures that the feature extractor learns representations that are both domain-invariant (due to $\mathcal{L}_{adv}$) and semantically discriminative across domains (due to $\mathcal{L}_{cont}$). Increasing these weights further (e.g., $\lambda_1 = 0.2, \lambda_2 = 0.3$ or $\lambda_1 = 0.3, \lambda_2 = 0.2$) leads to a slight decrease in accuracy. This suggests that excessively strong domain alignment constraints can potentially hinder the model's ability to learn fine-grained discriminative features for classification, leading to a trade-off. The sensitivity analysis thus

 doi:10.20944/preprints202512.2690.v1



confirms that careful tuning of these hyperparameters is essential for maximizing the generalization capabilities of **CD-SEAFNet**.

### *4.7. Human Evaluation of Semantic Understanding*

While quantitative metrics like accuracy are crucial, assessing the model's capacity for advanced semantic understanding and how it aligns with human perception provides deeper insights. We conducted a qualitative human evaluation study to compare how well **CD-SEAFNet** differentiates between semantically close or ambiguous actions, in contrast to a strong baseline model lacking explicit semantic enhancement.

#### 4.7.1. Methodology

A panel of 10 human participants was recruited to classify 200 skeleton sequences, specifically chosen to represent challenging pairs of actions (e.g., "walking" vs. "strolling," "punching" vs. "hitting lightly," "waving hand" vs. "clapping"). Each sequence was presented visually (as a short animation of joint movements). Participants were asked to classify each action and rate their confidence. We then compared human classification accuracy on these ambiguous pairs with:

1. **Base ST-GCN**: A representative state-of-the-art model without any explicit semantic enhancement or adaptive graph mechanisms.
2. **CD-SEAFNet (Full)**: Our complete proposed framework.

This evaluation aimed to understand if the semantic context fusion in **CD-SEAFNet** leads to classifications that are more consistent with human intuition, especially where visual cues alone might be insufficient or misleading.

#### 4.7.2. Results

The results of the human evaluation are summarized in Table 4.

**Table 4.** Performance on Ambiguous Action Pairs: Human vs. Models (% Accuracy)

| Method | Accuracy on Ambiguous Pairs |
|---|---|
| Human Evaluators | 96.1 |
| Base ST-GCN | 78.5 |
| **Ours (CD-SEAFNet)** | **91.3** |

Table 4 reveals that while human evaluators naturally achieved the highest accuracy on these semantically challenging action pairs, **CD-SEAFNet** significantly closed the gap compared to a strong **Base ST-GCN** model. The **Base ST-GCN** struggled more with subtle distinctions, indicating its reliance primarily on raw spatio-temporal patterns. In contrast, **CD-SEAFNet**'s performance of **91.3%** (a substantial improvement of 12.8% over the Base ST-GCN) suggests that the integration of high-level semantic descriptions enables the model to infer finer differences between actions, mimicking human cognitive processes more closely. This qualitative validation further reinforces the notion that explicit semantic enhancement, as implemented in **CD-SEAFNet**, is crucial for achieving a deeper and more robust understanding of human actions, moving beyond mere pattern recognition towards true semantic comprehension.

## 5. Conclusion

This paper tackled critical challenges in skeleton-based human action recognition, including domain shift, data noise, and complex action patterns, by introducing the Cross-Domain Semantic-Enhanced Adaptive Graph Fusion Network (CD-SEAFNet). CD-SEAFNet integrates three innovative modules: an Adaptive Spatio-Temporal Graph Feature Extractor for dynamic structure learning, a Semantic Context Encoder leveraging natural language for richer understanding, and a robust Domain Alignment Module for cross-domain generalization. Extensive experiments on NTU RGB+D datasets demonstrated CD-

 doi:10.20944/preprints202512.2690.v1



SEAFNet's superior performance, consistently outperforming state-of-the-art methods. Ablation studies validated the individual and synergistic contributions of its components, further supported by qualitative analysis showing enhanced distinction of semantically ambiguous actions. In conclusion, CD-SEAFNet represents a significant advancement by holistically combining adaptive graph learning, semantic context fusion, and robust domain alignment, leading to more intelligent and adaptable action recognition systems with a deeper, human-like understanding of actions.

## References


1. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
2. Chen, J.; Yang, D. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1380–1391. https://doi.org/10.18653/v1/2021.naacl-main.109.
3. Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N. Deduplicating Training Data Makes Language Models Better. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 8424–8445. https://doi.org/10.18653/v1/2022.acl-long.577.
4. Saxena, A.; Chakrabarti, S.; Talukdar, P. Question Answering Over Temporal Knowledge Graphs. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6663–6676. https://doi.org/10.18653/v1/2021.acl-long.520.
5. Zhou, Y. Sketch storytelling. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 4748–4752.
6. FitzGerald, J.; Hench, C.; Peris, C.; Mackie, S.; Rottmann, K.; Sanchez, A.; Nash, A.; Urbach, L.; Kakarala, V.; Singh, R.; et al. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 4277–4302. https://doi.org/10.18653/v1/2023.acl-long.235.
7. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.; Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 4148–4164. https://doi.org/10.18653/v1/2022.naacl-main.306.
8. Tian, Y.; Chen, G.; Song, Y. Aspect-based Sentiment Analysis with Type-aware Graph Convolutional Networks and Layer Ensemble. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2910–2922. https://doi.org/10.18653/v1/2021.naacl-main.231.
9. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed Acyclic Graph Network for Conversational Emotion Recognition. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1551–1560. https://doi.org/10.18653/v1/2021.acl-long.123.
10. Qin, L.; Wei, F.; Xie, T.; Xu, X.; Che, W.; Liu, T. GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 178–188. https://doi.org/10.18653/v1/2021.acl-long.15.
11. Fan, Z.; Gong, Y.; Liu, D.; Wei, Z.; Wang, S.; Jiao, J.; Duan, N.; Zhang, R.; Huang, X. Mask Attention Networks: Rethinking and Strengthen Transformer. In Proceedings of the Proceedings of the 2021 Conference of the

 doi:10.20944/preprints202512.2690.v1



North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1692–1701. https://doi.org/10.18653/v1/2021.naacl-main.135.

12. Cao, M.; Chen, L.; Shou, M.Z.; Zhang, C.; Zou, Y. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 9810–9823. https://doi.org/10.18653/v1/2021.emnlp-main.773.
13. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.
14. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 15757–15773. https://doi.org/10.18653/v1/2023.findings-emnlp.1055.
15. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
16. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
17. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 1610–1618. https://doi.org/10.18653/v1/2022.findings-acl.126.
18. Hardalov, M.; Arora, A.; Nakov, P.; Augenstein, I. Cross-Domain Label-Adaptive Stance Detection. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 9011–9028. https://doi.org/10.18653/v1/2021.emnlp-main.710.
19. Yu, T.; Liu, Z.; Fung, P. AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5892–5904. https://doi.org/10.18653/v1/2021.naacl-main.471.
20. Röttger, P.; Pierrehumbert, J. Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2400–2412. https://doi.org/10.18653/v1/2021.findings-emnlp.206.
21. Zhou, Y.; Zheng, X.; Hsieh, C.J.; Chang, K.W.; Huang, X. Defense against Synonym Substitution-based Adversarial Attacks via Dirichlet Neighborhood Ensemble. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5482–5492. https://doi.org/10.18653/v1/2021.acl-long.426.
22. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2035**.
23. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638274.
24. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2035**.