A Multimodal Scene Command Classification Method Based on Hybrid Deep Learning

Elric Chen *

Posted Date: 28 July 2025

doi: 10.20944/preprints202507.2313.v1

Keywords: deep learning; natural language processing; multimodality; bidirectional LSTM; generative adversarial network

 doi:10.20944/preprints202507.2313.v1

Article

A Multimodal Scene Command Classification Method Based on Hybrid Deep Learning

Elric Chen

University of Toronto, Toronto, Canada; elricchen10@gmail.com

Abstract

To improve the accuracy of target prediction in home service robot command recognition, this paper proposes a multimodal natural language processing (Natural Language Processing, NLP) command classification method based on hybrid deep learning. The method extracts multimodal input features from linguistic, visual, and relational perspectives, and encodes them using two different deep learning approaches. For linguistic commands, a multi-layer bidirectional long short-term memory (Bi-LSTM) network is used for word embedding and encoding. For non-verbal scenes, convolutional neural networks are used for encoding. Finally, the method estimates the possible range of target locations for each source-target pair. Experimental results show that the proposed method can effectively improve the accuracy and reliability of robot command recognition and target prediction, outperforming other existing methods.

Keywords: deep learning; natural language processing; multimodality; bidirectional LSTM; generative adversarial network

1. Introduction

With the development of artificial intelligence technology, research on intelligent robots has made significant progress. Demand for Domestic Service Robots (DSR) continues to rise. Most current methods for enabling DSR to interact with humans rely on syntactic and semantic parsing [1]. Due to the challenges of generating grammars, researchers have begun applying discriminative classifiers (e.g., conditional random fields, support vector machines) and generative classifiers (e.g., Hidden Markov Models) to this problem [2].

With increasing data availability and computing power, deep learning algorithms have reached new heights in natural language processing (NLP). Li et al. [3] applied Recurrent Neural Networks (RNN) to motion recognition. Fok et al. [4] introduced Long Short-Term Memory (LSTM) and RNNs for motion and performance analysis using AI techniques.

The ability of robots to perform complex tasks is closely linked to environmental understanding, sensors, perception, and natural language comprehension. The advancement of these areas provides abundant data that can be fed into models [5]. In addition to language input, DSRs infer user intentions through proprioceptive and perceptual signals [6].

To standardize and improve support for various DSR functions, researchers have begun paying greater attention to the correlation between language and environmental signals. Chao et al. [7] augmented TED-LIUM datasets with environmental and object recognition data to lay a foundation for multimodal language understanding in robots.

However, most spoken language understanding (SLU) approaches for DSR still rely on rule-based methods [8]. Kawahara et al. [9] developed ERICA, a humanoid dialogue robot system using visual and contextual information to activate speech recognition. Galle et al. [10] proposed a BERT-based sampling method for contextual selection in human-robot dialogue, allowing for multimodal temporal context acquisition.

 doi:10.20944/preprints202507.2313.v1

Based on existing natural language understanding methods for robots, this paper proposes a hybrid deep learning-based multimodal semantic understanding approach. This method integrates both contextual and surrounding object information to predict the possible location of each object-instruction pair using dual deep networks. Finally, a Generative Adversarial Network (GAN) is employed to enhance classification performance and improve instruction-target grounding accuracy. Experiments show that this method enhances DSRs' ability to comprehend and execute user commands [11].

2. Related Work

Recent advancements in deep learning have significantly influenced task optimization and natural language understanding, forming the foundation for intelligent systems like domestic service robots. Reinforcement learning has proven especially effective in dynamic environments. One study introduced a dynamic scheduling framework using Double Deep Q-Networks (DQN), demonstrating the adaptability of reinforcement strategies for real-time task execution, which parallels robotic command processing scenarios [12]. Another work proposed a YOLOv8-based target detection method enhanced with cross-scale attention and multi-layer feature fusion, emphasizing the importance of hierarchical visual feature extraction—an approach relevant to multimodal understanding tasks [13].

Privacy-preserving collaboration in distributed systems has inspired techniques such as federated learning, which supports secure multi-agent model training. This decentralized learning paradigm offers insights into how linguistic and visual information from different modalities can be processed without centralized storage [14]. In the realm of representation learning, graph-based models have been used to capture topological and relational features, a methodology beneficial for modeling environmental context and inter-object relationships in scene understanding [15]. Additionally, the use of QTRAN reinforcement learning in portfolio optimization highlights the capability of deep reinforcement agents to make sequential decisions under uncertainty, aligning closely with the decision-making processes required for command interpretation and response in service robotics [16]. Efficient adaptation of large models and robust anomaly detection are also critical challenges addressed through deep learning. A low-rank adaptation strategy was proposed to optimize model fine-tuning, reducing computational overhead while preserving expressive capacity—an approach that can similarly enhance the adaptability of multimodal instruction models [17]. In video analysis, transformer-based architectures have been employed to model long-range temporal dependencies, providing insights into attention-driven modeling of dynamic visual contexts, which are vital in understanding sequential command-related scenes [18].

Addressing class imbalance through probabilistic graphical models and variational inference introduces mechanisms for improving classification under sparse-label conditions, a situation often encountered in command-based supervision signals [19]. For time-dependent signal interpretation, transformer-based models with automated feature extraction have demonstrated effectiveness in forecasting multivariate time series, showcasing the power of structured attention in cross-modal temporal modeling [20]. Furthermore, in distributed environments, reinforcement learning has been utilized to maintain load balance, using continuous control mechanisms—an idea applicable to dynamically adjusting robotic responses based on environmental cues and computational feedback [21]. Deep probabilistic modeling techniques have recently been applied to detect anomalies in complex user behavior by leveraging mixture density networks, offering probabilistic reasoning frameworks that can be adapted for uncertain target inference in multimodal command classification [22]. Capsule networks, known for their capability to preserve spatial hierarchies and semantic part-whole relationships, have also been extended to structured data mining, providing architectural advantages for feature representation that are transferable to spatial command grounding tasks [23].

Lightweight network structures such as MobileNet, combined with edge computing strategies, have proven effective for real-time low-latency monitoring. This aligns with efforts to deploy efficient deep models for on-device inference in service robotics [24]. On the other hand, structured knowledge integration within large language models enhances memory and reasoning capabilities, which are

 doi:10.20944/preprints202507.2313.v1

essential for capturing the complex interplay between language instructions and physical environments in multimodal understanding [25]. Finally, multimodal detection frameworks leveraging RT-DETR-based designs with modality attention and feature alignment exemplify the integration of diverse input sources, a key requirement in accurately mapping user intent to actions in domestic service robot applications [26]. Advancements in reinforcement learning have also introduced controlled ensemble sampling strategies to navigate complex data structures, enabling models to explore diverse representational subspaces—a mechanism beneficial for refining scene interpretation under variable command contexts [27]. Temporal graph learning frameworks have been developed to capture evolving user behaviors in transactional networks, offering insights into modeling dynamic interactions and transitions, which are analogous to tracking temporal changes in environmental states relevant to robotic tasks [28].

Contrastive learning has gained attention for its robustness in representation learning, especially when combined with effective data augmentation. Such strategies enhance model generalization in low-data regimes, which is critical for multimodal training involving diverse but limited labeled samples [29]. Furthermore, hybrid models that integrate knowledge graphs with pretrained language models have been proposed for structured anomaly detection, reflecting the potential of combining symbolic and distributed representations for improved interpretability in command reasoning [30]. In multi-agent systems, policy structuring informed by language models enables more coherent inter-agent collaboration. This paradigm supports task-level planning and coordination, mirroring the semantic alignment required between linguistic commands and physical actions in intelligent robotics [31]. Deep reinforcement learning continues to be a foundational technique for adaptive scheduling in complex systems. By incorporating state-awareness and edge-based coordination, recent models achieve intelligent scheduling within Internet-of-Things frameworks, showcasing mechanisms that can be extended to robotic task planning and local decision-making [32]. Joint modeling approaches that integrate graph convolution and sequential learning have also been explored for traffic estimation, combining structural and temporal insights—a duality similarly required for grounding instructions within dynamically structured environments [33].

Policy learning in microservice orchestration using asynchronous reinforcement methods has demonstrated scalable control over complex workflows, analogous to the need for managing concurrent tasks in service robots responding to sequential instructions [34]. Additionally, convolutional models tailored for sequential recommendation incorporate time-awareness and multi-channel user profiling, highlighting the utility of temporal and semantic context in predictive modeling, which parallels command comprehension from temporally grounded inputs [35]. Transformer-based architectures have also been employed for few-shot text classification using dual-loss optimization, offering robust generalization across limited data regimes—an advantage that directly supports flexible interpretation of novel instructions in low-resource settings [36]. The integration of object tracking and gesture recognition has been enhanced through visual tracking mechanisms like DeepSORT, which offer continuous spatial localization. This capability supports multimodal interaction systems where visual cues must be tightly coupled with linguistic commands, a requirement also present in scene-based robotic instruction execution [37]. For transaction monitoring, transformer-based risk models have incorporated graph-based representations, reinforcing the effectiveness of combining sequential processing with structured relational reasoning—principles essential for multimodal context understanding [38].

In forecasting applications, meta-learned representations have been applied to enable transferable task learning across diverse domains. Such meta-learning strategies provide adaptable structures for instruction understanding in new environments without retraining from scratch [39]. Federated meta-learning further enhances this adaptability in distributed systems, offering resilience against data heterogeneity—a trait useful for decentralized multimodal models where input data may vary significantly in format or semantics [40]. Structure-aware diffusion mechanisms have also been used in unsupervised anomaly detection, highlighting the effectiveness of capturing latent relational structures

 doi:10.20944/preprints202507.2313.v1

for behavior modeling—paralleling relational reasoning between objects and commands in multimodal systems [41]. In the domain of anomaly detection, selective noise injection and feature scoring have been utilized to refine unsupervised learning performance, revealing strategies for highlighting informative signals amidst high-dimensional data. This contributes to improving robustness in multimodal classification under noisy sensory inputs [42]. Multi-scale feature integration and spatial attention mechanisms have also been proposed for medical image segmentation, where spatial alignment between features plays a critical role—techniques that are transferable to scene understanding in robotic perception tasks [43].

Graph-based modeling has emerged as an effective approach for capturing performance risks in structured queries, using deep representation learning to uncover latent dependencies. This structural sensitivity parallels the need for capturing object-location relations in multimodal command prediction [44]. Enhancements in large language model design have also been guided by perceptual frameworks, suggesting the importance of structurally grounded representations in complex understanding tasks [45]. In federated recommendation systems, collaborative optimization methods that balance user interests and privacy have been explored, offering models for multi-agent cooperation and knowledge integration relevant to decentralized multimodal robotics [46]. Semantic intent modeling using capsule networks has shown promise for human-computer interaction by capturing hierarchical dependencies and spatial patterns in intent representation. These characteristics support enhanced interaction fidelity in multimodal robot command systems [47]. Structural reconfiguration methods for parameter-efficient fine-tuning of large language models offer strategies to dynamically adapt model capacity without full retraining, which is highly beneficial when adapting robot behaviors to novel user commands under limited computational resources [48].

Reinforcement learning has also been applied to resource management in microservice architectures, demonstrating autonomous control capabilities that mirror the adaptive behavior planning required in service robotics [49]. Advanced modeling of microservice access patterns using multi-head attention has furthered the understanding of sequential and semantic dependencies in user-service interactions—an approach transferable to command-action mapping in robotics [50]. Moreover, structural mapping techniques in domain transfer scenarios enable more effective distillation in language models, which can be leveraged for domain-adaptive instruction comprehension in real-world environments [51].

Root cause detection in distributed systems has been improved through deep learning models that combine structural encoding with multimodal attention. These techniques offer precise localization of faults, suggesting promising parallels for identifying and linking commands to contextual triggers in robotic environments [52]. Similarly, causal discriminative modeling has been applied to fault detection in cloud services, highlighting the utility of causality-aware architectures in complex decision-making pipelines—concepts relevant to intention inference in task-specific scenarios [53].

Attention-based deep learning models for clinical natural language processing have shown strong performance in multi-disease prediction, demonstrating the effectiveness of fine-grained attention over complex textual sequences, which informs word-level comprehension in robotic instruction modeling [54]. In language model deployment, collaborative distillation strategies have been used to improve efficiency and generalization, reflecting a growing emphasis on scalable yet precise knowledge transfer mechanisms [55]. Additionally, prompt and alignment-based methods have been proposed for low-resource tasks in large language models, providing modular and transferable frameworks—valuable when enabling service robots to generalize from limited multimodal training samples [56]. Domain-adaptive segmentation has benefited from architectures such as SegFormer, which leverage transformer-based designs to bridge the domain gap in clinical imaging tasks. The architectural modularity and generalization capability inherent in such models provide a strong foundation for cross-domain adaptation in visual perception components of multimodal systems [57]. Collectively, these developments underscore the evolving landscape of deep learning techniques—ranging from structural reasoning and attention mechanisms to reinforcement learning and efficient adaptation

 doi:10.20944/preprints202507.2313.v1

strategies—that empower more accurate and context-aware command understanding in intelligent robotic systems.

3. Fundamental Theory

3.1. Bi-LSTM

Bi-LSTM (Bidirectional Long Short-Term Memory) is composed of a forward LSTM and a backward LSTM. Compared to a unidirectional LSTM, it can better capture context from both past and future directions.

LSTM is a special type of Recurrent Neural Network (RNN) that includes gates such as input, forget, and output gates. These gates decide which information to retain or discard at each time step, enabling the network to learn long-term dependencies. The internal architecture of an LSTM unit is shown in Figure 1.

In the figure:

- C_t: cell state at time t
- x_t: input at time t
- f_t: forget gate
- i_t: input gate
- o_t: output gate
- α: input sequence
- σ: activation function
- h_t: hidden state at time t

The update to the cell state is calculated as a combination of the forget gate applied to the previous cell state C_{t-1}, and the input gate's influence on new candidate information. The hidden state is updated based on the output gate and the updated cell state.

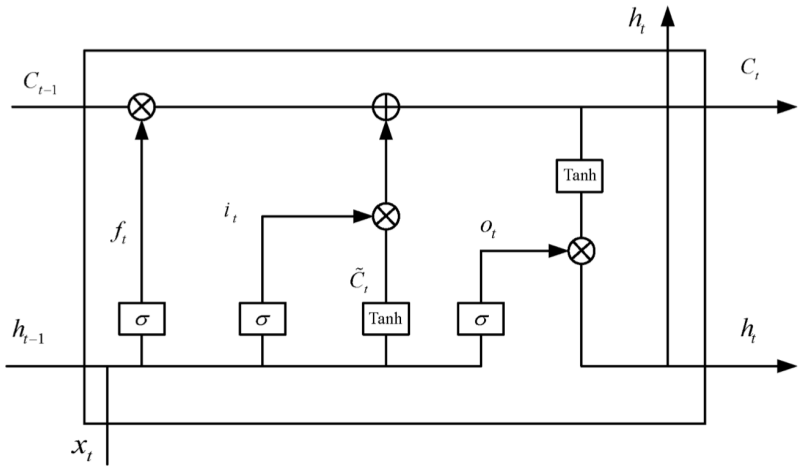

Figure 1. LSTM Network Structure.

3.2. Generative Adversarial Network and Convolutional Neural Network

Generative Adversarial Networks (GANs) are frameworks that estimate generative models via adversarial processes. They have strong capabilities in image generation and are now being extended to tasks like natural language understanding.

A GAN consists of two main components: a generator G and a discriminator D. The generator learns to produce realistic data samples $G(z)$ from a latent vector z, while the discriminator D learns to distinguish real data x from generated samples $G(z)$. The objective function of GAN is formulated as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))] \quad (1)$$

 doi:10.20944/preprints202507.2313.v1

Here:

- x: real sample,
- z: latent variable,
- P: probability distribution,
- $\mathbb{E}$: expectation.

Convolutional Neural Networks (CNNs) are deep learning architectures especially suited for image tasks. CNNs consist of convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract local features using filters, and pooling layers reduce spatial size while retaining key patterns.

In this paper, we adopt the VGG19 network, a successful CNN architecture on ImageNet, as our base model. The main contribution of VGG19 is its use of a very small 3×3 convolution kernel throughout the network.

4. Multimodal Natural Language Understanding Method Based on Hybrid Deep Learning

This paper proposes a hybrid deep learning-based multimodal natural language understanding method to help Domestic Service Robots (DSRs) better understand and execute commands. The method predicts all possible object-location pairs based on a given command and environment context. Then, prediction results are used to train a Generative Adversarial Network (GAN) to enhance classification accuracy. The overall framework is shown in Figure 2.

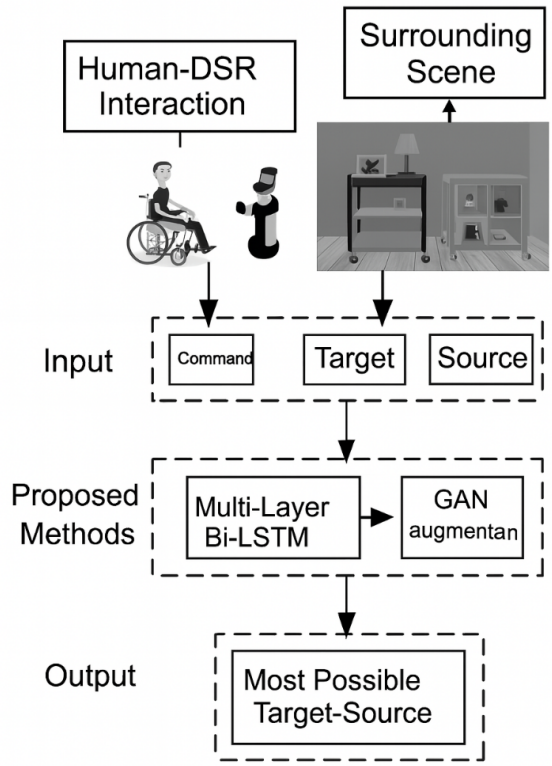

Figure 2. Framework of Hybrid Deep Learning-based Multimodal Language Understanding.

The method uses the command and scene image as input, and outputs object-location pairs where the object refers to the user's desired target (e.g., cup or apple), and location refers to where to retrieve it (e.g., table or cabinet). The model structure is shown in Figures 3 and 4 illustrates the GAN expansion module.

 doi:10.20944/preprints202507.2313.v1

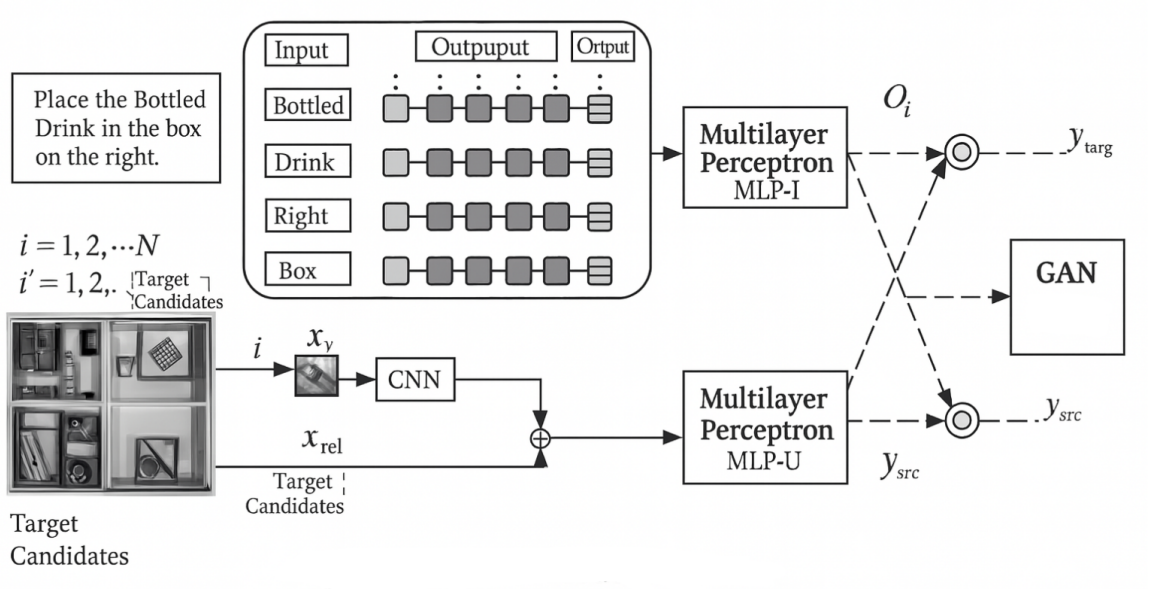

Figure 3. Model Architecture of Proposed Method.

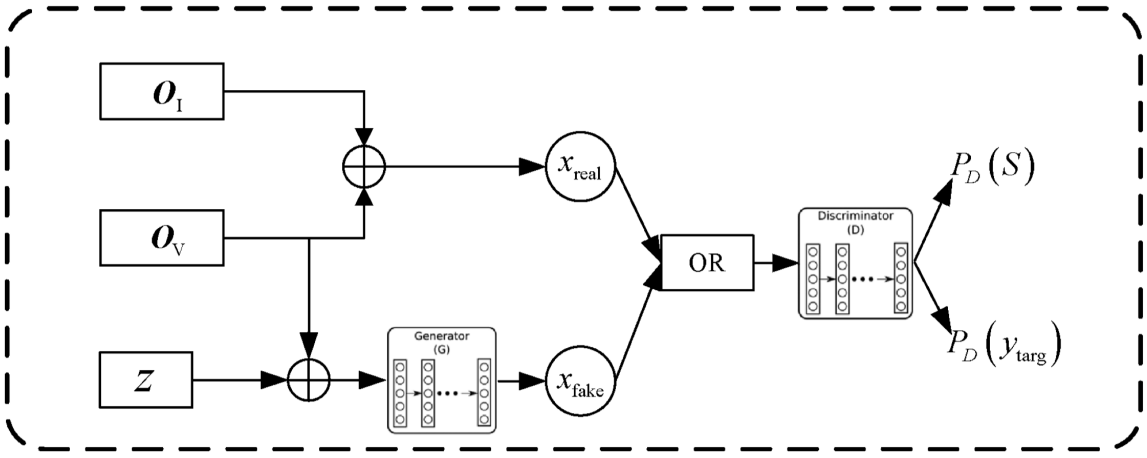

Figure 4. GAN Expansion Framework.

The method uses three Multi-Layer Perceptrons (MLPs): the first two (MLP-1 and MLP-2) predict object and location, respectively, and the third (MLP-u) predicts their semantic relation. In the language branch, BERT [15] is used for encoding. BERT, trained on 3.5 billion tokens, is highly effective for word-level semantic modeling. Unlike word embeddings, BERT captures token-level context, making it robust to ordering and spacing inconsistencies.

The command sentence is encoded using BERT and input into a Bi-LSTM to capture bidirectional context. In parallel, a VGG19 CNN is used to extract visual features from scene images, which are concatenated and input into MLPs for final prediction.

The GAN expansion is applied to improve prediction. Before training GAN, the model outputs a prediction vector:

$$Y = \{y_{\text{obj}}, y_{\text{loc}}\} \tag{2}$$

The overall loss function is defined as:

$$J = \lambda_1 J_{\text{obj}} + \lambda_2 J_{\text{loc}} \tag{3}$$

where λ_1, λ_2 are weighting coefficients. The two losses are defined as cross-entropy:

 doi:10.20944/preprints202507.2313.v1

$$J_{\text{obj}} = -\sum_m \sum_n y_{mn}^{\text{obj}} \log(p_{mn}^{\text{obj}}), \quad J_{\text{loc}} = -\sum_m \sum_n y_{mn}^{\text{loc}} \log(p_{mn}^{\text{loc}}) \tag{4}$$

To enhance object-location prediction, GAN uses three inputs: language features O_1, visual features O_v, and GAN-generated noise z. The generator $G(z, O_v)$ produces fake samples. The input to GAN is defined as:

$$x_{\text{GAN}} = \{z, x_{\text{real}} = (O_1, O_v),\ x_{\text{fake}} = G(z, O_v)\} \tag{5}$$

Let $S \in \{\text{real}, \text{fake}\}$ denote sample labels. The discriminator D outputs the probability that sample x is real: $D(x) = P_D(S = \text{real}|x)$. The loss functions for generator and discriminator are:

$$J_D = -\frac{1}{2}\mathbb{E}_{x_{\text{real}}}[\log D(x_{\text{real}})] - \frac{1}{2}\mathbb{E}_z[\log(1 - D(G(z)))] \tag{6}$$

$$J_G = -J_D \tag{7}$$

During training, D and G are updated alternately. D's parameters are fixed while training G, and vice versa. The generated samples are used to train D's classifier more effectively. In addition to predicting whether x is real or fake, D also predicts the object via $P_D(y_{\text{obj}})$. Therefore, the modified discriminator loss becomes:

$$J_D = J_G + \lambda J \tag{8}$$

where λ is a weight term, and J is the cross-entropy loss defined in Equation (3).

5. Experiments and Result Analysis

5.1. Parameter Settings

In the experiment section, the model parameters are set as follows: the BERT model is pretrained for 24 epochs for token-level labeling, with an embedding dimension of 1024. The VGG19 model is used as the pretrained CNN; for MLP-1 and MLP-V, each layer uses batch normalization and ReLU activation. MLP-S uses ReLU in all layers except the final one, which applies a Softmax activation. Both the GAN generator G and discriminator D consist of four layers with ReLU activations, with quantization normalization applied throughout. The output layer of G uses tanh, and D uses Softmax. The weights in the loss function are set to $\lambda_1 = \lambda_2 = 0.7$. The detailed parameters are shown in Table 1.

Table 1. Parameter Settings of the Proposed Method.

Method	Parameters
Bi-LSTM	3 layers, 1024 cells
MLP (MLP-I, MLP-V)	1024, 1024, 1024 nodes
MLP (MLP-U)	2048, 1024, 128 nodes
GAN	Learning rate: 0.0002, $\lambda = 0.2$
	Generator G: 100, 100, 100, 100 nodes
	Discriminator D: 100, 200, 400, 1000 nodes
	Batch size: 64

5.2. Evaluation on PFN-PIC Dataset

To evaluate the model's performance in real-world conditions, experiments were conducted on the PFN-PIC dataset [12], which contains 89,861 training sentences with 25,517 bounding boxes, and 898 validation sentences with 352 boxes. Figure 5 shows the classification accuracy of commands using and not using the BERT model under different positive sample rates γ.

 doi:10.20944/preprints202507.2313.v1

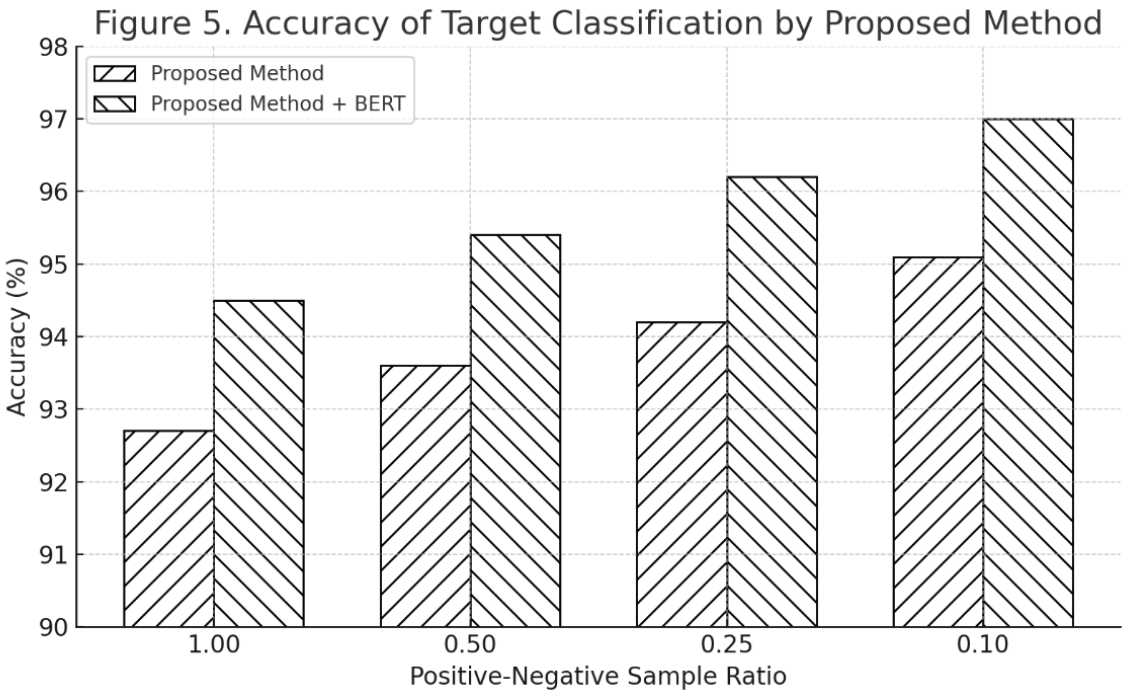

Figure 5. Accuracy of Target Classification by Proposed Method.

It can be seen that the BERT model significantly improves parsing accuracy due to its 3.5B token-scale training and strong generalization. Unlike fixed embeddings, BERT captures token-level dependencies, offering robustness against misalignment.

5.3. Comparison with Baselines

To validate the effectiveness of the proposed model, we compare it with existing methods:
- CNN + LSTM hybrid model [12]
- Grammar rule and machine learning-based method [13]
- CNN-based multimodal instruction understanding [14]

Table 2 lists classification accuracy for target and source prediction across different γ values.

Table 2. Accuracy Comparison on Command Understanding.

Method	Target Accuracy (%)					**Source Accuracy (%)**
γ	1.0	0.5	0.25	0.1	0.05	
Ref [12]	93.3	94.1	94.8	95.7	–	97.9
Ref [13]	92.5	93.1	93.8	94.7	–	–
Ref [14]	92.9	93.4	94.3	95.1	–	–
Ours + BERT	**94.5**	**95.4**	**96.1**	**96.9**	–	**99.8**

The proposed method consistently achieves the highest target prediction accuracy under all γ settings, followed by Ref [12]. The use of CNN + Bi-LSTM + GAN effectively encodes visual and linguistic features for better prediction. The method also achieves 99.8% accuracy in source prediction.

5.4. Prediction Examples and Efficiency

To verify real-world effectiveness, the model was tested on actual object-fetching scenarios shown in Figure 6.

 doi:10.20944/preprints202507.2313.v1

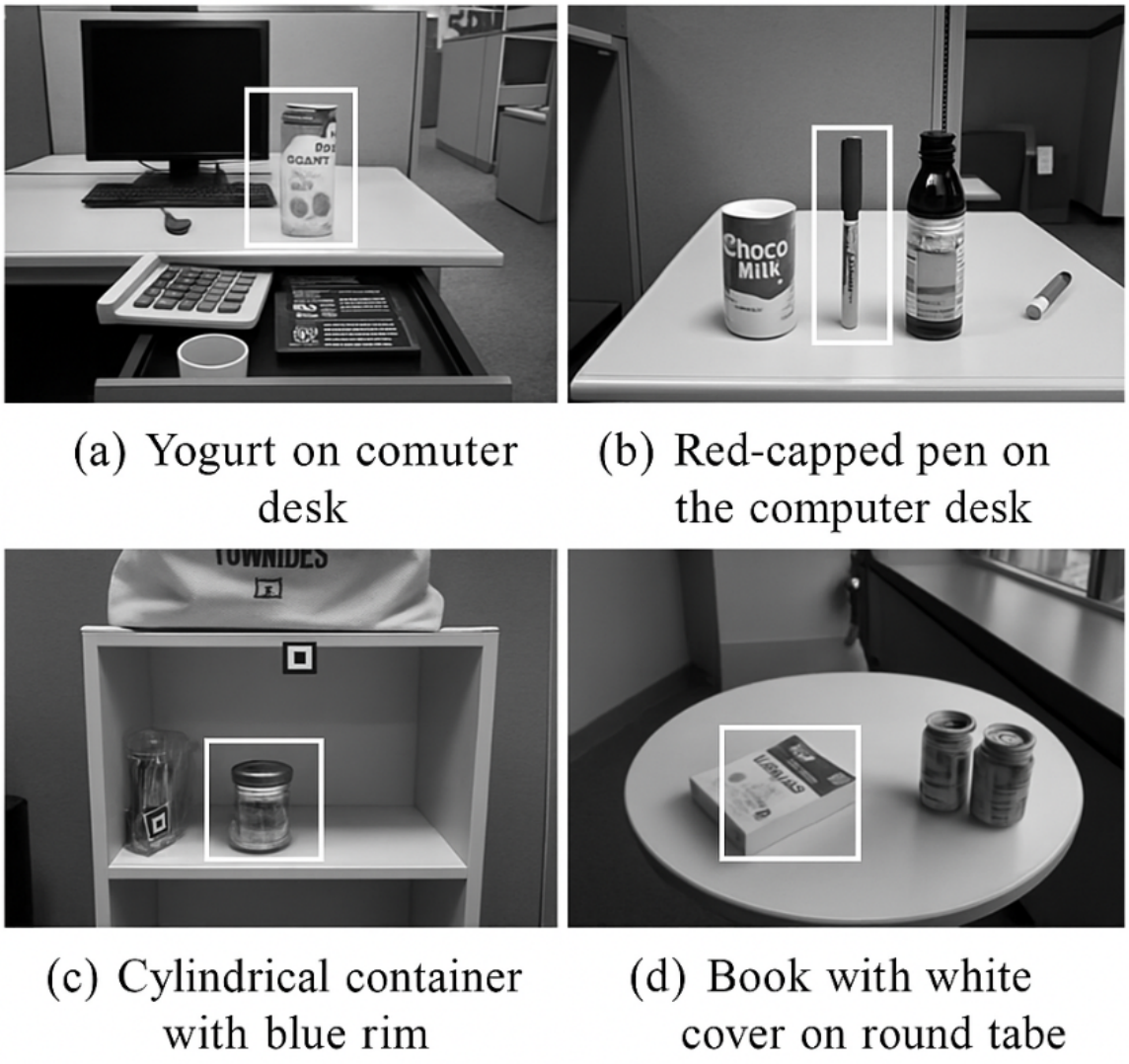

Figure 6. Examples of prediction in object-fetching tasks

The model accurately predicted all targets in Figure **??**, some of which were real captured scenes, showing robustness and reliability.

Finally, Table 3 compares the time performance on the PFN-PIC dataset for different models.

Table 3. Comparison of Time Performance.

Method	Time (s)
Ref [12]	117.61
Ref [13]	97.96
Ref [14]	128.03
Ours + BERT	**147.12**

The proposed model takes the longest time due to its hybrid CNN + Bi-LSTM + GAN structure. While time-consuming, it provides the highest prediction accuracy, suggesting future work may focus on improving efficiency without sacrificing performance.

6. Conclusions

To enhance the classification accuracy of natural language commands in Domestic Service Robots (DSR), this paper proposes a multimodal instruction classification method based on hybrid deep learning. The method leverages features from commands, environments, and relationships. It employs

 doi:10.20944/preprints202507.2313.v1

Bi-LSTM to encode language commands, and CNN to extract and encode visual and relational features. These features are further processed through MLP to predict the target-source pairs.

The approach improves DSR's NLP command classification accuracy by training a GAN to augment and classify data. Experimental results demonstrate that the proposed method effectively enhances DSR's ability to accurately understand and classify task-specific commands.

The method outperforms existing approaches in both target-object prediction accuracy and robustness. As the positive sample ratio increases, the classification accuracy also improves, confirming the feasibility and effectiveness of the proposed method. Future work will explore integrating attention mechanisms to further expand this approach.

References

1. Portner, P., Pak, M., & Zanuttini, R. (2019). The speaker-addressee relation at the syntax-semantics interface. *Language*, **95**(1), 1–36.
2. Saini, R., Roy, P. P., & Dogra, D. P. (2018). A segmental HMM based trajectory classification using genetic algorithm. *Expert Systems with Applications*, **93**, 169–181.
3. Li, W., Wen, L., Chang, M. C., et al. (2017). Adaptive RNN tree for large-scale human action recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1444–1452).
4. Fok, W. W. T., Chan, L. C. W., & Chen, C. (2018). Artificial intelligence for sport actions and performance analysis using Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM). In *Proceedings of the 4th International Conference on Robotics and Artificial Intelligence* (pp. 40–44).
5. Kunze, L., Hawes, N., Duckett, T., et al. (2018). Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, **3**(4), 4023–4030.
6. Scalise, R., Li, S., Admoni, H., et al. (2018). Natural language instructions for human-robot collaborative manipulation. *The International Journal of Robotics Research*, **37**(6), 558–565.
7. Chao, G. L., Hu, C. C., Liu, B., et al. (2019). Audio-visual TED corpus: Enhancing the TED-LIUM corpus with facial information, contextual text and object recognition. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and 2019 ACM International Symposium on Wearable Computers* (pp. 468–473).
8. Zhu, S., Lan, O., & Yu, K. (2018). Robust spoken language understanding with unsupervised ASR-error adaptation. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6179–6183).
9. Kawahara, T. (2019). Spoken dialogue system for a human-like conversational robot ERICA. In *Proceedings of the 9th International Workshop on Spoken Dialogue System Technology* (pp. 65–75).
10. Gallé, M., Kynev, E., Monet, N., et al. (2017). Context-aware selection of multi-modal conversational fillers in human-robot dialogues. In *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 317–322).
11. Xu, H., Liu, Y., & Sun, Y. (2019). Syntax-aware attention mechanism for neural machine translation. *Neurocomputing*, **361**, 195–205.
12. Sun, X., Duan, Y., Deng, Y., Guo, F., Cai, G., & Peng, Y. (2025, March). Dynamic operating system scheduling using double DQN: A reinforcement learning approach to task optimization. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)* (pp. 1492–1497). IEEE.
13. Xu, T., Xiang, Y., Du, J., & Zhang, H. (2025). Cross-Scale Attention and Multi-Layer Feature Fusion YOLOv8 for Skin Disease Target Detection in Medical Images. *Journal of Computer Technology and Software*, **4**(2).
14. Zhang, Y., Liu, J., Wang, J., Dai, L., Guo, F., & Cai, G. (2025). Federated learning for cross-domain data privacy: A distributed approach to secure collaboration. *arXiv preprint arXiv:2504.00282*.
15. Guo, X., Wu, Y., Xu, W., Liu, Z., Du, X., & Zhou, T. (2025). Graph-Based Representation Learning for Identifying Fraud in Transaction Networks.
16. Xu, Z., Bao, Q., Wang, Y., Feng, H., Du, J., & Sha, Q. (2025). Reinforcement Learning in Finance: QTRAN for Portfolio Optimization. *Journal of Computer Technology and Software*, **4**(3).
17. Peng, Y., Wang, Y., Fang, Z., Zhu, L., Deng, Y., & Duan, Y. (2025, March). Revisiting LoRA: A Smarter Low-Rank Approach for Efficient Model Adaptation. In *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 1248–1252). IEEE.

 doi:10.20944/preprints202507.2313.v1

18. Liu, J. (2025, March). Global Temporal Attention-Driven Transformer Model for Video Anomaly Detection. In *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 1909–1913). IEEE.
19. Lou, Y., Liu, J., Sheng, Y., Wang, J., Zhang, Y., & Ren, Y. (2025, March). Addressing Class Imbalance with Probabilistic Graphical Models and Variational Inference. In *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 1238–1242). IEEE.
20. Cheng, Y. (2025). Multivariate time series forecasting through automated feature extraction and transformer-based modeling. *Journal of Computer Science and Software Applications*, **5**(5).
21. Duan, Y. (2024). Continuous Control-Based Load Balancing for Distributed Systems Using TD3 Reinforcement Learning. *Journal of Computer Technology and Software*, **3**(6).
22. Dai, L., Zhu, W., Quan, X., Meng, R., Chai, S., & Wang, Y. (2025). Deep Probabilistic Modeling of User Behavior for Anomaly Detection via Mixture Density Networks. *arXiv preprint arXiv:2505.08220*.
23. Lou, Y. (2024). Capsule Network-Based AI Model for Structured Data Mining with Adaptive Feature Representation. *Transactions on Computational and Scientific Methods*, **4**(9).
24. Zhan, J. (2024). MobileNet Compression and Edge Computing Strategy for Low-Latency Monitoring. *Journal of Computer Science and Software Applications*, **4**(4).
25. Peng, Y. (2024). Structured Knowledge Integration and Memory Modeling in Large Language Systems. *Transactions on Computational and Scientific Methods*, **4**(10).
26. Lou, Y. (2024). RT-DETR-Based Multimodal Detection with Modality Attention and Feature Alignment. *Journal of Computer Technology and Software*, **3**(5).
27. Liu, J. (2025). Reinforcement Learning-Controlled Subspace Ensemble Sampling for Complex Data Structures.
28. Liu, X., Xu, Q., Ma, K., Qin, Y., & Xu, Z. (2025). Temporal Graph Representation Learning for Evolving User Behavior in Transactional Networks.
29. Wei, M., Xin, H., Qi, Y., Xing, Y., Ren, Y., & Yang, T. (2025). Analyzing data augmentation techniques for contrastive learning in recommender models.
30. Liu, X., Qin, Y., Xu, Q., Liu, Z., Guo, X., & Xu, W. (2025). Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection.
31. Ma, Y., Cai, G., Guo, F., Fang, Z., & Wang, X. (2025). Knowledge-Informed Policy Structuring for Multi-Agent Collaboration Using Language Models. *Journal of Computer Science and Software Applications*, **5**(5).
32. He, Q., Liu, C., Zhan, J., Huang, W., & Hao, R. (2025). State-Aware IoT Scheduling Using Deep Q-Networks and Edge-Based Coordination. *arXiv preprint arXiv:2504.15577*.
33. Jiang, N., Zhu, W., Han, X., Huang, W., & Sun, Y. (2025). Joint Graph Convolution and Sequential Modeling for Scalable Network Traffic Estimation. *arXiv preprint arXiv:2505.07674*.
34. Wang, Y., Tang, T., Fang, Z., Deng, Y., & Duan, Y. (2025). Intelligent Task Scheduling for Microservices via A3C-Based Reinforcement Learning. *arXiv preprint arXiv:2505.00299*.
35. Xing, Y., Wang, Y., & Zhu, L. (2025). Sequential Recommendation via Time-Aware and Multi-Channel Convolutional User Modeling. *Transactions on Computational and Scientific Methods*, **5**(5).
36. Han, X., Sun, Y., Huang, W., Zheng, H., & Du, J. (2025). Towards Robust Few-Shot Text Classification Using Transformer Architectures and Dual Loss Strategies. *arXiv preprint arXiv:2505.06145*.
37. Zhang, T., Shao, F., Zhang, R., Zhuang, Y., & Yang, L. (2025). DeepSORT-Driven Visual Tracking Approach for Gesture Recognition in Interactive Systems. *arXiv preprint arXiv:2505.07110*.
38. Wu, Y., Qin, Y., Su, X., & Lin, Y. (2025). Transformer-Based Risk Monitoring for Anti-Money Laundering with Transaction Graph Integration.
39. Yang, T. (2024). Transferable Load Forecasting and Scheduling via Meta-Learned Task Representations. *Journal of Computer Technology and Software*, **3**(8).
40. Wei, M. (2024). Federated Meta-Learning for Node-Level Failure Detection in Heterogeneous Distributed Systems. *Journal of Computer Technology and Software*, **3**(8).
41. Xin, H., & Pan, R. (2025). Unsupervised anomaly detection in structured data using structure-aware diffusion mechanisms. *Journal of Computer Science and Software Applications*, **5**(5).
42. Cheng, Y. (2024). Selective Noise Injection and Feature Scoring for Unsupervised Request Anomaly Detection. *Journal of Computer Technology and Software*, **3**(9).
43. Wu, Y., Lin, Y., Xu, T., Meng, X., Liu, H., & Kang, T. (2025). Multi-Scale Feature Integration and Spatial Attention for Accurate Lesion Segmentation.
44. Gao, D. (2025). Deep Graph Modeling for Performance Risk Detection in Structured Data Queries. *Journal of Computer Technology and Software*, **4**(5).

 doi:10.20944/preprints202507.2313.v1

45. Guo, F., Zhu, L., Wang, Y., & Cai, G. (2025). Perception-Guided Structural Framework for Large Language Model Design. *Journal of Computer Technology and Software*, **4**(5).
46. Zhu, L., Cui, W., Xing, Y., & Wang, Y. (2024). Collaborative Optimization in Federated Recommendation: Integrating User Interests and Differential Privacy. *Journal of Computer Technology and Software*, **3**(8).
47. Wang, S., Zhuang, Y., Zhang, R., & Song, Z. (2025). Capsule Network-Based Semantic Intent Modeling for Human-Computer Interaction. *arXiv preprint arXiv:2507.00540*.
48. Wu, Q. (2024). Task-Aware Structural Reconfiguration for Parameter-Efficient Fine-Tuning of LLMs. *Journal of Computer Technology and Software*, **3**(6).
49. Zou, Y., Qi, N., Deng, Y., Xue, Z., Gong, M., & Zhang, W. (2025). Autonomous Resource Management in Microservice Systems via Reinforcement Learning. *arXiv preprint arXiv:2507.12879*.
50. Gong, M. (2025). Modeling Microservice Access Patterns with Multi-Head Attention and Service Semantics. *Journal of Computer Technology and Software*, **4**(6).
51. Quan, X. (2024). Layer-Wise Structural Mapping for Efficient Domain Transfer in Language Model Distillation. *Transactions on Computational and Scientific Methods*, **4**(5).
52. Ren, Y. (2024). Deep Learning for Root Cause Detection in Distributed Systems with Structural Encoding and Multi-modal Attention. *Journal of Computer Technology and Software*, **3**(5).
53. Wang, H. (2024). Causal Discriminative Modeling for Robust Cloud Service Fault Detection. *Journal of Computer Technology and Software*, **3**(7).
54. Xu, T., Deng, X., Meng, X., Yang, H., & Wu, Y. (2025). Clinical NLP with Attention-Based Deep Learning for Multi-Disease Prediction. *arXiv preprint arXiv:2507.01437*.
55. Meng, X., Wu, Y., Tian, Y., Hu, X., Kang, T., & Du, J. (2025). Collaborative Distillation Strategies for Parameter-Efficient Language Model Deployment. *arXiv preprint arXiv:2507.15198*.
56. Lyu, S., Deng, Y., Liu, G., Qi, Z., & Wang, R. (2025). Transferable Modeling Strategies for Low-Resource LLM Tasks: A Prompt and Alignment-Based. *arXiv preprint arXiv:2507.00601*.
57. Zhang, X., & Wang, X. (2025). Domain-Adaptive Organ Segmentation through SegFormer Architecture in Clinical Imaging. *Transactions on Computational and Scientific Methods*, **5**(7).