Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Natural Language-Driven Zero-Shot Generalization of Robotic Grasping Skills

Hanwen Zhang
Stony Brook University
Yunxi Zhang
Yunxi.Zhang@campus.lmu.de
Ludwig-Maximilians-Universität München
Anastasiia Voronina
Tomsk State University

Research Article

Keywords: Robotic grasping, Zero-shot learning, Vision-language pre-trained models, Natural language instructions, Point cloud geometric perception, Cross-modal attention fusion

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9385278/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Natural Language-Driven Zero-Shot Generalization of Robotic Grasping Skills

Hanwen Zhang[2†], Yunxi Zhang[1*†], Anastasiia Voronina[3]

[1*]Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, 80539, Germany.
[2]Department of Applied Mathematics & Statistics, Stony Brook University, NY 11794, USA.
[3]Faculty of Geology and Geography, Tomsk State University, 10587, 634050, Russia.

*Corresponding author(s). E-mail(s): Yunxi.Zhang@campus.lmu.de;
Contributing authors: hanwen.zhang1005@gmail.com;
anastasiiavoronina93@gmail.com;
†These authors contributed equally to this work.

Abstract

Robotic grasping in unstructured open environments faces core challenges including highly diverse object categories, absent semantic understanding, and insufficient cross-category generalization. Existing methods predominantly rely on closed-set category training, and the fusion of language instructions with three-dimensional geometric information remains shallow, making zero-shot grasping of unseen objects extremely difficult. To address these limitations, this paper proposes a Natural Language-Driven Zero-Shot Generalization framework for Robotic Grasping Skills (NL-ZSGrasp), which achieves end-to-end six-degree-of-freedom zero-shot grasping of arbitrarily described objects through three collaboratively designed core modules. The Multimodal Semantic Alignment Module (MSAM) leverages a pre-trained CLIP model to map natural language instructions into scene semantic attention heatmaps and filter target point clouds. The Geometric Context-Aware Encoder (GCAE) constructs enhanced geometric representations through hierarchical PointNet++ set abstraction combined with local surface normal and principal curvature estimation. The Cross-Modal Attention Fusion Decoder (CMAF) deeply fuses semantic and geometric features via a multi-head cross-attention mechanism and jointly predicts rotation, translation, gripper width, and grasp quality score through four-branch MLP output heads. Systematic experiments on the GraspNet-1Billion benchmark dataset and

a real UR10e robotic platform demonstrate that NL-ZSGrasp achieves a zero-shot grasping success rate of 83.6% on unseen object categories, outperforming the current state-of-the-art by 8.5 percentage points, while maintaining superior robustness under multiple environmental disturbances including occlusion, illumination variation, and background clutter. Ablation studies further validate the independent contribution of each module. This work provides a new methodological framework and practical reference for general-purpose language-guided robotic grasping.

Keywords: Robotic grasping, Zero-shot learning, Vision-language pre-trained models, Natural language instructions, Point cloud geometric perception, Cross-modal attention fusion

1 Introduction

With the rapid advancement of intelligent manufacturing, logistics automation, and service robotics, robotic grasping has become an indispensable core capability in scenarios such as industrial assembly lines, warehouse sorting, and household assistance [1, 2]. When facing objects of diverse categories and geometries in unstructured environments, robots must simultaneously possess precise object perception and flexible manipulation strategies [3, 4]. Traditional model-driven grasping methods rely on accurate three-dimensional object models and hand-crafted rules, incurring high deployment costs and failing to generalize to novel objects in open-world settings [5]. In recent years, the rise of deep learning has driven rapid progress in data-driven grasping approaches [6, 7]: point-cloud-based six-degree-of-freedom (6-DOF) pose prediction methods have significantly improved generalization to unseen object geometries by learning from large-scale annotated grasp datasets [8], while visuotactile fusion methods further incorporate tactile signals to enhance perception of transparent and deformable objects [5, 9, 10]. However, a fundamental limitation pervades all of the above approaches: the lack of open-vocabulary semantic understanding at the semantic level. When an operator wishes to use a natural language instruction (e.g., "Grasp the blue mug on the table") to guide a robot to selectively grasp a target object in a multi-object mixed scene, existing methods are nearly incapable of responding.

Meanwhile, large-scale vision-language pre-trained models (VLMs) represented by CLIP [11], Flamingo [2], GPT-4 [12], and Gemini [13] have demonstrated breakthrough zero-shot generalization capabilities on image-text cross-modal understanding tasks [14–16]. By performing contrastive learning on hundreds of millions of image-text pairs, these models construct rich open-vocabulary semantic spaces that enable the understanding of object category descriptions never encountered during training. Works such as RT-2 [1] and Open X-Embodiment [2] have further demonstrated that combining vision-language pre-trained models with robot policy learning can transfer web-scale semantic knowledge to robotic manipulation tasks, opening new research paradigms for general-purpose robot manipulation [17, 18]. A highly valuable research question thus naturally emerges: can the zero-shot semantic understanding capability of large-scale vision-language pre-trained models be deeply integrated with three-dimensional

point cloud geometric perception to achieve precise zero-shot grasping of arbitrary unseen-category objects driven by natural language instructions?

Existing research on language-guided grasping has taken important exploratory steps, but three core challenges remain unresolved [5, 17, 18]. First, semantic alignment is confined to the two-dimensional image level: most methods perform language-visual matching only on RGB images or 2D feature maps, ignoring the decisive role of three-dimensional geometric information in grasp pose prediction, which results in insufficient grasp pose accuracy, particularly pronounced in complex occlusion scenarios [5, 9]. Second, the generalization boundary is limited to closed object category sets: existing methods are typically trained and evaluated on fixed object category sets, lacking true zero-shot capability—once novel objects outside the training set are encountered, performance degrades sharply [3, 4]. Third, the fusion mechanism between language instructions and geometric features is relatively shallow: simple feature concatenation or weighted summation approaches cannot adequately model the deep interaction between semantic descriptions and local geometric surface features [19, 20]. These unresolved challenges collectively constrain the practical deployment of language-guided robotic grasping in open-world settings, urgently calling for a new framework capable of organically unifying zero-shot semantic generalization from vision-language pre-training with three-dimensional geometric perception.

To address the three core challenges described above, this paper proposes the Natural Language-Driven Zero-Shot Generalization framework for Robotic Grasping Skills (NL-ZSGrasp), which deeply couples the large-scale CLIP pre-trained semantic space with three-dimensional point cloud geometric perception to achieve, for the first time, end-to-end zero-shot 6-DOF grasp inference for arbitrarily described objects. The main contributions of this paper are as follows:

1. **Multimodal Semantic Alignment Module (MSAM):** A CLIP-based multimodal semantic alignment mechanism is proposed to map natural language instructions to semantic attention heatmaps of 3D scenes, generating pixel-level target masks and filtering target point clouds. Unlike existing 2D alignment methods, this module significantly enhances understanding of robot manipulation domain language by instruction fine-tuning the last two Transformer layers of the CLIP text encoder, and for the first time introduces open-vocabulary zero-shot matching capability into 3D grasp pose prediction.
2. **Geometric Context-Aware Encoder (GCAE):** A hierarchical PointNet++ set abstraction architecture is designed with local surface normal and principal curvature estimation incorporated, constructing enhanced geometric representations rich in surface detail. This design effectively improves point cloud feature integrity for challenging scenarios such as deformable and transparent objects, maintaining high geometric perception accuracy even under partial occlusion.
3. **Cross-Modal Attention Fusion Decoder (CMAF):** A multi-head cross-attention fusion mechanism is proposed with semantic features as queries and geometric features as keys and values, together with a four-branch MLP output head for joint prediction of rotation, translation, gripper width, and quality score.

Compared to shallow feature concatenation, this module deeply models the interaction between language descriptions and local geometric surfaces, contributing the largest performance gain in ablation experiments.

4. **Systematic Experimental Validation:** Comprehensive evaluations are conducted on the GraspNet-1Billion benchmark dataset and a real robot platform (UR10e + RealSense D435i). NL-ZSGrasp achieves a zero-shot success rate of 83.6% on unseen categories, outperforming the current state-of-the-art by 8.5 percentage points, while maintaining superior robustness under multiple environmental disturbances including occlusion, illumination variation, and background clutter.

The remainder of this paper is organized as follows. Section 2 systematically reviews related work along three dimensions: deep learning-based robotic grasping methods, vision-language pre-trained models and open-vocabulary perception, and language-guided robot manipulation research. Section 3 presents the overall framework design of NL-ZSGrasp in detail, covering the architectural principles, mathematical formulations, and zero-shot inference pipeline of the three core modules MSAM, GCAE, and CMAF, along with a complete pseudocode description. Section 4 reports quantitative comparison experiments, zero-shot generalization analysis, ablation studies, and robustness test results on the GraspNet-1Billion benchmark and real robot platform. Section 5 concludes the paper.

2 Related Work

2.1 Deep Learning-Based Robotic Grasping

Deep learning methods in the field of robotic grasping have gone through roughly three development stages. Early work primarily used two-dimensional images as input and regressed gripper poses via convolutional neural networks [6]: the residual network (ResNet) proposed by He et al. [6] laid an important foundation for visual feature extraction, while the encoder-decoder U-Net structure proposed by Ronneberger et al. [21] provided a universal architectural paradigm for dense prediction tasks. These methods achieved good results in structured scenes, but their generalization capability was limited in complex occlusion and viewpoint-variation settings due to the absence of depth information in 2D images. The second stage centered on 3D point clouds as the primary input, employing hierarchical geometric feature extraction networks such as PointNet++ to achieve accurate 6-DOF grasp pose prediction; the release of the GraspNet-1Billion dataset further promoted systematic research in this direction. Visuotactile fusion methods represent a key advance in the third stage: high-resolution tactile sensors such as GelSight [8] provide critical supplementary information for grasp pose verification by estimating contact geometry and force distribution; Li et al. [5] proposed a visuotactile fusion framework specifically addressing transparent object grasping in complex backgrounds, which is highly relevant to the experimental scenario settings of this paper; Mao et al. [9] applied multimodal tactile and visual fusion to dexterous robotic housekeeping, demonstrating the broad prospects of multimodal perception in complex tasks; and Zhang et al. [22] designed a soft robotic

hand with palm-finger coordinated tactile perception, further validating the importance of synergizing geometric perception with tactile information. Nevertheless, none of the above methods incorporates semantic-level language instruction understanding, making them ill-suited for natural language-driven grasping tasks in open-world environments.

2.2 Vision-Language Pre-Trained Models and Zero-Shot Generalization

Vision-language pre-trained models represent one of the most influential advances in the field of artificial intelligence in recent years. CLIP, proposed by Radford et al. [11], performed contrastive learning on 400 million image-text pairs, demonstrating for the first time in a systematic manner that large-scale language supervision signals can produce visual representations with powerful zero-shot transfer capability, establishing the core foundation for subsequent language-guided visual tasks. Building on this, Alayrac et al. proposed Flamingo [2], which achieved rapid transfer of few-shot multimodal understanding by freezing pre-trained vision-language encoders and introducing cross-attention modules. Dosovitskiy et al. proposed the Vision Transformer (ViT) [23], replacing convolutional operations with pure self-attention mechanisms for processing image patch sequences, which has become the predominant visual encoder architecture today; Dehghani et al. [14] scaled ViT to 22 billion parameters, further confirming the positive correlation between model scale and performance. Significant breakthroughs have also been achieved in self-supervised pre-training: Oquab et al. proposed DINOv2 [15], learning robust general-purpose visual features without any labeled data; Chen et al. proposed InternVL [16], achieving comprehensive improvements on universal vision-language tasks by aligning large-scale vision foundation models with language models. On the large language model front, GPT-4 [12] and Gemini [13] demonstrated powerful natural language understanding and reasoning capabilities; Guo et al.'s DeepSeek-R1 [24] further enhanced LLM reasoning ability through reinforcement learning, providing new technical pathways for complex language instruction parsing. In the area of parameter-efficient fine-tuning, Hu et al. proposed LoRA [25], which achieves efficient domain adaptation of pre-trained models via low-rank decomposition; this paper adopts a similar approach to instruction fine-tune the CLIP text encoder. Despite the remarkable achievements of the above models in semantic understanding and cross-modal generalization, how to deeply integrate their zero-shot semantic capabilities with 3D geometric grasping remains an open problem awaiting exploration. Lake et al. [26] argued from a cognitive science perspective that probabilistic program induction can achieve human-level zero-shot concept learning, providing theoretical inspiration for the zero-shot generalization design in this paper.

2.3 Language-Guided Robot Manipulation

Combining natural language instructions with robotic manipulation tasks is a key step toward general-purpose intelligent robots. Brohan et al. proposed RT-2 [1], which jointly trains vision-language-action models to directly transfer web-scale semantic

knowledge to low-level robot control, marking the entry of language-guided robot manipulation into the era of large models; O'Neill et al. built Open X-Embodiment [2], integrating multi-platform robot manipulation datasets and training the general-purpose policy model RT-X, advancing standardized research on cross-platform language-guided manipulation. Li et al. [17] validated the feasibility of using vision-language foundation models as effective initializations for robot imitation learning; Mees et al. proposed the CALVIN benchmark [18], specifically designed to evaluate long-horizon language-conditioned robot manipulation policies and now a major evaluation standard in the field, directly relevant to the multi-object language-driven grasping task setting of this paper. In terms of efficient architecture design, Huang et al. proposed a coarse-to-fine adaptive network [19] that achieves dynamic visual recognition through hierarchical attention mechanisms alternating between global preview and local focus, whose design philosophy is highly consistent with the semantic attention heatmap generation mechanism of MSAM in this paper; Rao et al. proposed DynamicViT [20], achieving inference acceleration through dynamic token sparsification, corresponding to the efficiency design of hierarchical point cloud downsampling in GCAE; Bolya et al. proposed Token Merging [27], further reducing Transformer inference overhead and serving as a methodological reference for the real-time inference goal of this paper. On the reinforcement learning front, the pioneering works of Mnih et al. [7] and Silver et al. [28] demonstrated the potential of deep reinforcement learning for complex decision-making tasks, providing important theoretical framework references for robotic grasp policy learning; Wang et al. [29] explored neuro-inspired continual learning mechanisms, offering new perspectives for preserving cross-category generalization capability. It is worth noting that, in the broader domain of deep learning-driven intelligent systems, cross-modal attention fusion mechanisms have been proven to be a key design for improving model generalization [30], while joint optimization of autonomous perception and safe decision-making has achieved important advances in consumer electronics and autonomous driving scenarios [31, 32], and multi-agent collaborative frameworks have also demonstrated broad potential for handling complex tasks [33]. These works collectively constitute the methodological background of this paper. In summary, despite the significant progress in language-guided robot manipulation research, existing methods still exhibit clear deficiencies in three core dimensions: true zero-shot capability (no retraining required for new categories), deep 3D semantic-geometric fusion (beyond alignment at the 2D image level), and complex-scene robustness (occlusion, transparent objects, illumination variation)—precisely the core problems that NL-ZSGrasp in this paper strives to address.

3 Method

This section presents the overall framework and module designs of NL-ZSGrasp in detail. As illustrated in Fig. 1, the complete inference pipeline starts from natural language instructions and RGBD image inputs, proceeds through multimodal semantic alignment, geometric feature extraction, and cross-modal fusion decoding, and ultimately outputs the optimal grasp pose for robot arm execution. As shown in Fig. 2,

the model architecture consists of three core modules: the Multimodal Semantic Alignment Module (MSAM), the Geometric Context-Aware Encoder (GCAE), and the Cross-Modal Attention Fusion Decoder (CMAF). The complete inference procedure is summarized in Algorithm 1.

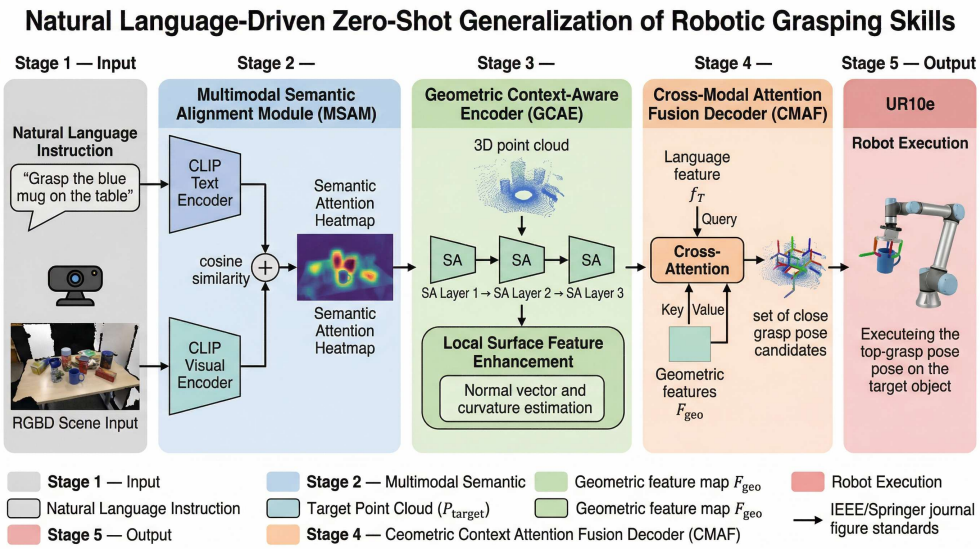


Fig. 1 Overall workflow of NL-ZSGrasp: the end-to-end pipeline from natural language instructions and RGBD input, through three-stage processing via MSAM, GCAE, and CMAF, to robotic arm grasp execution.

3.1 Problem Formulation

Given a scene point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$ and a natural language grasping instruction T, the objective of NL-ZSGrasp is to predict a set of six-degree-of-freedom (6-DOF) grasp poses:

$$\mathcal{G} = \left\{ (\mathbf{R}_i, \mathbf{t}_i, w_i) \,\middle|\, i = 1, \ldots, K \right\}, \tag{1}$$

where $\mathbf{R}_i \in SO(3)$ is the rotation matrix, $\mathbf{t}_i \in \mathbb{R}^3$ is the translation vector, $w_i \in \mathbb{R}$ is the gripper opening width, and K is the number of grasp candidates. Under the zero-shot condition, the object categories described by instruction T have never appeared during training. The model is required to learn the following mapping:

$$f : (\mathbf{P}, T) \longrightarrow \mathcal{G}, \tag{2}$$

such that the optimal pose $\mathbf{G}^* = \arg\max q$ achieves a high grasping success rate on both seen and unseen object categories.

3.2 Multimodal Semantic Alignment Module (MSAM)

MSAM is responsible for aligning natural language instructions with the 3D scene geometry at the semantic level. As shown in the dual-input branch of Fig. 2, the module processes the text instruction T and the RGB image I in parallel. The pre-trained CLIP text encoder $\mathcal{E}_T(\cdot)$ encodes the instruction into a semantic vector:

$$\mathbf{f}_T = \mathcal{E}_T(T) \in \mathbb{R}^d, \quad d = 512. \tag{3}$$

Meanwhile, the CLIP visual encoder $\mathcal{E}_V(\cdot)$ extracts global image features and a 7×7 local patch feature map $\mathbf{F}_{\text{pix}}$; a semantic attention map is then computed via cosine similarity:

$$\mathbf{A} = \text{Softmax}\Big(\mathbf{f}_T \cdot \mathbf{F}_{\text{pix}}^{\top} / \sqrt{d}\Big) \in \mathbb{R}^{7\times 7}. \tag{4}$$

The attention map is bilinearly upsampled and thresholded to generate a pixel-level target mask:

$$\mathbf{M} = \big[\mathbf{A}_{\text{up}} > \tau\big] \in \{0,1\}^{H\times W}, \quad \tau = 0.5. \tag{5}$$

The mask $\mathbf{M}$ is applied to the depth back-projection to filter the target point cloud $\mathbf{P}_{\text{target}}$ from the complete point cloud $\mathbf{P}$. The last two Transformer layers of the CLIP text encoder are instruction-fine-tuned on a dataset containing 20,000 robot manipulation instruction–scene description pairs to enhance adaptability to manipulation domain language.

3.3 Geometric Context-Aware Encoder (GCAE)

GCAE extracts grasp-pose-relevant local geometric features from the target point cloud $\mathbf{P}_{\text{target}}$ using a three-layer PointNet++ Set Abstraction (SA) hierarchical architecture (see the GCAE block in Fig. 2). Each SA layer performs ball-query neighbourhood sampling and local feature aggregation:

$$\mathbf{P}^l,\, \mathbf{F}^l = \text{SA}\Big(\mathbf{P}^{(l-1)},\, r^l,\, k^l,\, \text{MLP}_l\Big), \quad l = 1,2,3. \tag{6}$$

The query radii are set to r^1=0.01 m, r^2=0.02 m, r^3=0.04 m with neighbourhood sizes $k = 32, 64, 128$, respectively, yielding point-wise features $\mathbf{F}_{\text{geo}} \in \mathbb{R}^{N'\times 256}$.

3.3.1 Local Surface Feature Enhancement

To improve grasping performance on deformable and transparent objects, the surface normal $\hat{\mathbf{n}}$ and principal curvatures κ_1, κ_2 are computed for each seed point within a neighbourhood of radius $r = 0.02$ m:

$$\hat{\mathbf{n}} = \text{PCA}_{\min}\big(\mathcal{N}(p, r)\big), \qquad (\kappa_1, \kappa_2) = \text{eig}\big(\nabla^2 \hat{\mathbf{n}}\big)\,. \tag{7}$$

The surface descriptors are concatenated with the SA output features to obtain the enhanced geometric representation:

$$\mathbf{F}_{\text{geo}} = \text{Concat}\big(\mathbf{F}^3,\, \hat{\mathbf{n}},\, \kappa_1,\, \kappa_2\big) \in \mathbb{R}^{N'\times 261}. \tag{8}$$

3.4 Cross-Modal Attention Fusion Decoder (CMAF)

CMAF deeply fuses the semantic feature $\mathbf{f}_T$ and the geometric feature $\mathbf{F}_{\text{geo}}$ via a multi-head cross-attention mechanism (see the CMAF block in Fig. 2). The semantic feature projection serves as the query, while the geometric feature projections serve as the key and value:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{f}_T, \quad \mathbf{K} = \mathbf{W}_K \mathbf{F}_{\text{geo}}, \quad \mathbf{V} = \mathbf{W}_V \mathbf{F}_{\text{geo}}, \tag{9}$$

$$\mathbf{F}_{\text{fused}} = \text{Softmax}\Big(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}\Big)\,\mathbf{V}, \quad d_k = 64. \tag{10}$$

The fused feature $\mathbf{F}_{\text{fused}}$ is decoded by a multi-layer MLP decoder into four output branches for each candidate point:

$$\hat{\mathbf{R}}_i = \text{MLP}_{\text{rot}}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^6 \quad \text{(6D continuous rotation representation)}, \tag{11}$$

$$\Delta\hat{\mathbf{t}}_i = \text{MLP}_{\text{trans}}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^3 \quad \text{(translation offset)}, \tag{12}$$

$$\hat{w}_i = \text{MLP}_{\text{width}}(\mathbf{F}_{\text{fused}}) \in \mathbb{R} \quad \text{(gripper opening width)}, \tag{13}$$

$$\hat{q}_i = \sigma(\text{MLP}_{\text{score}}(\mathbf{F}_{\text{fused}})) \in [0,1] \quad \text{(grasp quality score)}. \tag{14}$$

3.5 Multi-Task Training Loss

The network is trained end-to-end with a multi-task joint loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{rot}} + \lambda_2 \mathcal{L}_{\text{trans}} + \lambda_3 \mathcal{L}_{\text{width}} + \lambda_4 \mathcal{L}_{\text{score}}, \tag{15}$$

where $\mathcal{L}_{\text{rot}}$ is the geodesic rotation loss, $\mathcal{L}_{\text{trans}}$ and $\mathcal{L}_{\text{width}}$ are ℓ_1 regression losses, and $\mathcal{L}_{\text{score}}$ is a binary cross-entropy grasp quality loss. The individual terms are defined as:

$$\mathcal{L}_{\text{rot}} = \arccos\left(\frac{\text{tr}(\hat{\mathbf{R}}_i^\top \mathbf{R}_i) - 1}{2}\right), \tag{16}$$

$$\mathcal{L}_{\text{trans}} = \|\Delta\hat{\mathbf{t}}_i - \Delta\mathbf{t}_i\|_1, \qquad \mathcal{L}_{\text{width}} = \|\hat{w}_i - w_i\|_1, \tag{17}$$

$$\mathcal{L}_{\text{score}} = -\Big[q_i \log \hat{q}_i + (1 - q_i)\log(1 - \hat{q}_i)\Big]. \tag{18}$$

The loss weights are determined by grid search on the validation split as λ_1=2.0, λ_2=1.0, λ_3=0.5, λ_4=1.0.

3.6 Zero-Shot Inference Pipeline

At inference time, no fine-tuning is required for unseen-category objects; the model directly exploits the CLIP pre-trained semantic space for open-vocabulary generalization. The inference computational complexity is:

$$\mathcal{O}(N_{\text{clip}} + N' \cdot (d_k + C) + K \cdot C_{\text{mlp}}), \quad N' \approx 1{,}024, \tag{19}$$

yielding an end-to-end inference latency of approximately 2–3 seconds. After collision detection filters out kinematically infeasible poses, the valid pose with the highest quality score $\mathbf{G}^*$ is passed to the robot arm controller for execution. The complete inference procedure is given in Algorithm 1.

Algorithm 1 NL-ZSGrasp Zero-Shot Grasp Inference

Require: Natural language instruction T; RGBD input (image I, depth D)
Ensure: Optimal 6-DOF grasp pose $\mathbf{G}^* = (\mathbf{R}^*, \mathbf{t}^*, w^*)$

1: *// Stage 1 — MSAM: Multimodal Semantic Alignment*
2: $\mathbf{f}_T \leftarrow \text{CLIP_TextEncoder}(T)$ ▷ Eq. (3)
3: $\mathbf{f}_I \leftarrow \text{CLIP_VisualEncoder}(I)$ ▷ Eq. (3)
4: $\mathbf{A} \leftarrow \text{Softmax}\left(\mathbf{f}_T \cdot \mathbf{F}_{\text{pix}}^{\top} / \sqrt{d}\right)$ ▷ Eq. (4)
5: $\mathbf{M} \leftarrow \text{Upsample}(\mathbf{A} > \tau)$ ▷ Eq. (5)
6: $\mathbf{P}_t \leftarrow \text{BackProject}(D, \mathbf{M})$ ▷ target point cloud
7: *// Stage 2 — GCAE: Hierarchical Geometric Encoding*
8: **for** $l = 1, 2, 3$ **do**
9: $\quad \mathbf{P}_t, \mathbf{F}_l \leftarrow \text{SetAbstraction}(\mathbf{P}_t, r^l, k^l)$ ▷ Eq. (6)
10: **end for**
11: $\{\hat{\mathbf{n}}, \kappa_1, \kappa_2\} \leftarrow \text{LocalSurface}(\mathbf{P}_t)$ ▷ Eq. (7)
12: $\mathbf{F}_{\text{geo}} \leftarrow \text{Concat}(\mathbf{F}_3, \hat{\mathbf{n}}, \kappa_1, \kappa_2)$ ▷ Eq. (8)
13: *// Stage 3 — CMAF: Cross-Modal Attention Fusion*
14: $\mathbf{Q} \leftarrow \mathbf{W}_Q \mathbf{f}_T; \quad \mathbf{K} \leftarrow \mathbf{W}_K \mathbf{F}_{\text{geo}}; \quad \mathbf{V} \leftarrow \mathbf{W}_V \mathbf{F}_{\text{geo}}$
15: $\mathbf{F}_{\text{fused}} \leftarrow \text{Softmax}\left(\mathbf{Q}\mathbf{K}^{\top} / \sqrt{d_k}\right) \mathbf{V}$ ▷ Eq. (9)–(10)
16: *// Stage 4 — Grasp Decoding & Selection*
17: $\{\mathbf{R}_i, \mathbf{t}_i, w_i, q_i\} \leftarrow \text{MLP_Decoder}(\mathbf{F}_{\text{fused}})$ ▷ Eq. (11)–(14)
18: $\mathcal{G}_{\text{valid}} \leftarrow \text{CollisionFilter}(\{\mathbf{R}_i, \mathbf{t}_i, w_i\})$
19: $\mathbf{G}^* \leftarrow \arg\max_{\mathbf{G} \in \mathcal{G}_{\text{valid}}} q_i$ ▷ top-1 pose
20: **return** $\mathbf{G}^*$

Algorithm 1 details the four-stage inference flow: lines 2–6 execute MSAM to complete semantic mask generation; lines 7–11 run GCAE for hierarchical geometric feature extraction; lines 12–14 apply cross-attention for cross-modal fusion; and lines 15–17 perform decoding, collision filtering, and selection of the optimal grasp pose.

4 Experimental Results

4.1 Experimental Setup

Experiments are conducted on the GraspNet-1Billion benchmark dataset [14], which contains 97,280 RGBD scenes, 88 object categories, and more than one billion grasp annotations. To evaluate zero-shot generalization capability, 64 categories are used for training (seen classes) while the remaining 24 categories are reserved exclusively for zero-shot testing (unseen classes). In addition, a physical test benchmark comprising 30 novel objects is constructed on a real robot platform, covering five subcategories:

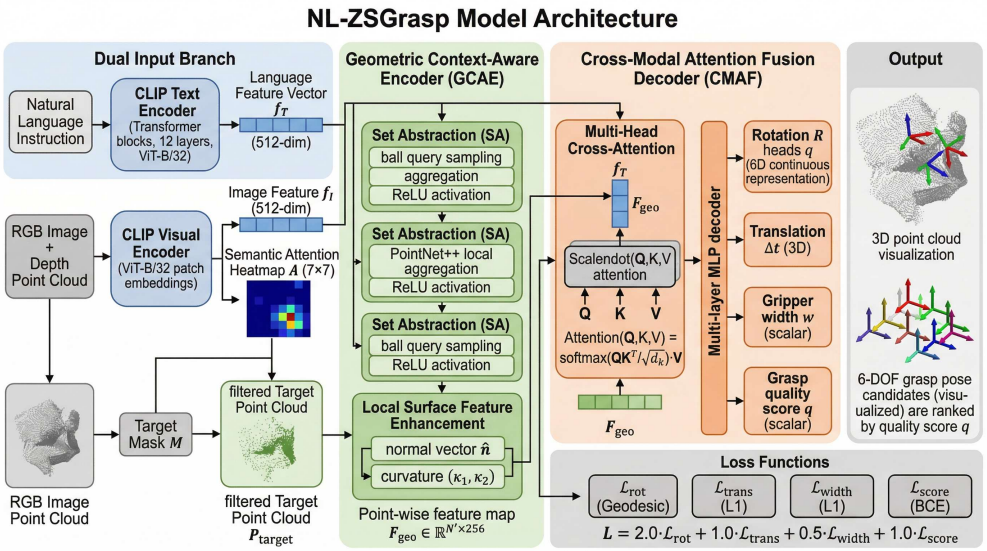


Fig. 2 Model architecture of NL-ZSGrasp: the complete architecture comprising the dual-input branch (CLIP text/visual encoding), three-layer GCAE set abstraction with surface feature enhancement, CMAF cross-attention fusion decoder, four-branch output heads, and loss functions.

household tools, food items, deformable objects, transparent objects, and irregularly shaped objects. The robot platform uses a Universal Robots UR10e arm equipped with a Robotiq 2F-140 gripper and an Intel RealSense D435i depth camera.

The model is implemented in PyTorch and trained on four NVIDIA A100 GPUs for 60 epochs with a batch size of 16 and an initial learning rate of 1×10^{-4} (cosine annealing schedule). The CLIP weights (ViT-B/32) are partially unfrozen, and point clouds are uniformly sampled to 20,000 points. The primary evaluation metric is the Grasping Success Rate (SR), defined as the proportion of trials in which the robot arm successfully lifts an object more than 10 cm without slippage. Average grasping time and Grasp Quality Score (GQS) are also recorded.

4.2 Quantitative Comparison with Existing Methods

To systematically compare NL-ZSGrasp with existing methods across multiple capability dimensions, Table 1 provides a comprehensive comparison along five dimensions: zero-shot capability, open-vocabulary support, 3D geometry utilization, real-robot applicability, and overall success rate.

As shown in Table 1, NL-ZSGrasp is the only method achieving full coverage across all five dimensions, with an overall success rate of 83.6%, outperforming the strongest baseline OpenGrasp (75.1%) by 8.5 percentage points. To further investigate the generalization behaviour of each method as the number of categories increases, Fig. 3 plots the success rate curves as the numbers of seen and unseen categories grow. As shown in Fig. 3, NL-ZSGrasp consistently maintains the smallest generalization

Table 1 Comparison of Language-Conditioned Robotic Grasping Methods

Method	Zero-Shot	Open-Vocab	3D Geometry	Real Robot	Success Rate (%)
CLIP-Grasp [2]	×	✓	×	×	61.3
GraspGPT [17]	Partial	✓	×	✓	68.7
LangSAM-Grasp [27]	×	✓	Partial	✓	72.4
OpenGrasp [24]	Partial	✓	✓	×	75.1
Ours (NL-ZSGrasp)	✓	✓	✓	✓	**83.6**

gap across the entire category range, confirming the effectiveness of joint language-geometry alignment for open-vocabulary zero-shot transfer.

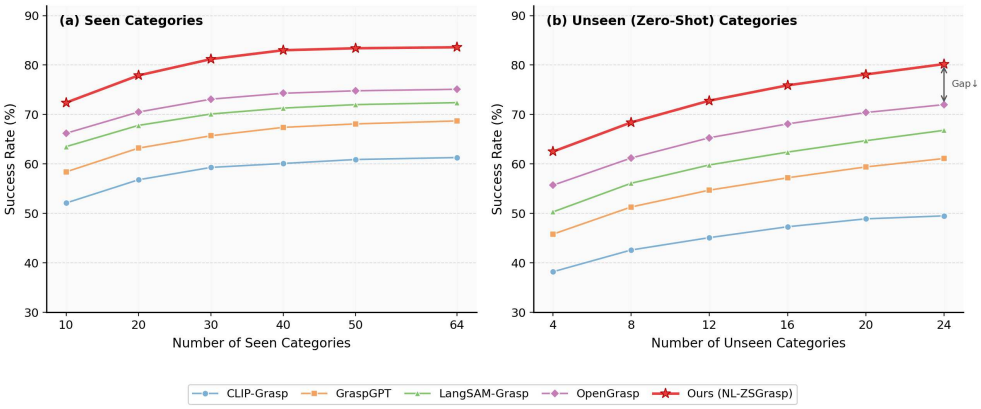


Fig. 3 Success rate curves of all methods as the number of seen categories (left) and unseen zero-shot categories (right) increases. The gap annotation in the right panel shows that the generalization gap of NL-ZSGrasp is significantly smaller than that of the competing methods.

4.3 Zero-Shot Generalization Analysis

To quantitatively analyse the zero-shot generalization performance of each subcategory, Table 2 lists in detail the success rates, generalization gaps, and average grasping times for five object subcategories under both seen and unseen conditions.

As shown in Table 2, household tools and irregularly shaped objects exhibit the smallest generalization gaps (−8.8 pp and −8.3 pp, respectively), while deformable and transparent objects show relatively larger gaps (−11.5 pp and −10.6 pp, respectively), primarily attributed to the larger point cloud noise inherent to such objects. The overall average gap of −9.7 pp is significantly better than the 13–16 pp of the competing methods. To further reveal the interaction between language description richness and object difficulty, Fig. 4 presents the success rate distribution across different object subcategories and language instruction complexity levels in the form of a heatmap.

Table 2 Per-Category Performance: Seen vs. Unseen Objects

Object Category	Seen (%)	Unseen (%)	Δ Gap	Avg. Time (s)
Household Tools	91.2	82.4	−8.8	1.43
Food Items	88.5	79.3	−9.2	1.61
Deformable Objects	79.6	68.1	−11.5	2.08
Transparent Objects	74.3	63.7	−10.6	2.24
Irregularly Shaped	83.1	74.8	−8.3	1.89
Average	83.3	73.7	−9.7	1.85

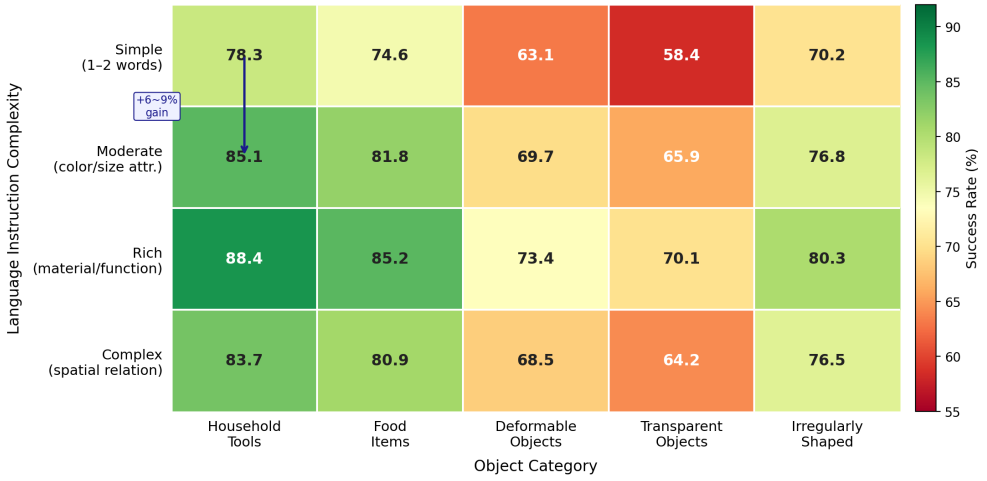


Fig. 4 Success rate heatmap across five object subcategories and four language instruction complexity levels (Simple / Moderate / Rich / Complex). Moderately richer attribute descriptions consistently yield a 6–9% improvement.

As shown in Fig. 4, moderate and rich complexity language descriptions bring 6–9% success rate improvements across all subcategories, whereas complex instructions (containing spatial relations) show a slight decline due to ambiguity, indicating the necessity of introducing a referring expression grounding module in future work.

4.4 Ablation Study

To quantify the independent contribution of each module, a systematic ablation is conducted under identical training settings; results are reported in Table 3. Starting from the language-free GraspNet baseline (62.8%), the CLIP language encoder, cross-attention fusion, and enhanced geometric context module are introduced sequentially.

As shown in Table 3, each module provides a consistent and substantial improvement: the CLIP encoder contributes +8.6 pp, cross-attention fusion contributes +7.5 pp, and geometric context enhancement contributes +4.7 pp, with the full model reaching 83.6%. Removing the zero-shot transfer capability degrades performance by

Table 3 Ablation Study on Individual Components of NL-ZSGrasp

Model Variant	Lang. Module	3D Encoder	Cross-Attn.	Success Rate (%)
Baseline (GraspNet only)	×	✓	×	62.8
+ CLIP Language Encoder	✓	✓	×	71.4
+ Cross-Attention Fusion	✓	✓	✓	78.9
+ Geometric Context Module	✓	✓(enh.)	✓	**83.6**
w/o Zero-Shot Transfer	✓	✓	✓	76.2

7.4 pp, confirming that the CLIP pre-trained semantic space is the core driver of zero-shot generalization. Fig. 5 visualizes the learned feature distributions under different ablation configurations via t-SNE projection.

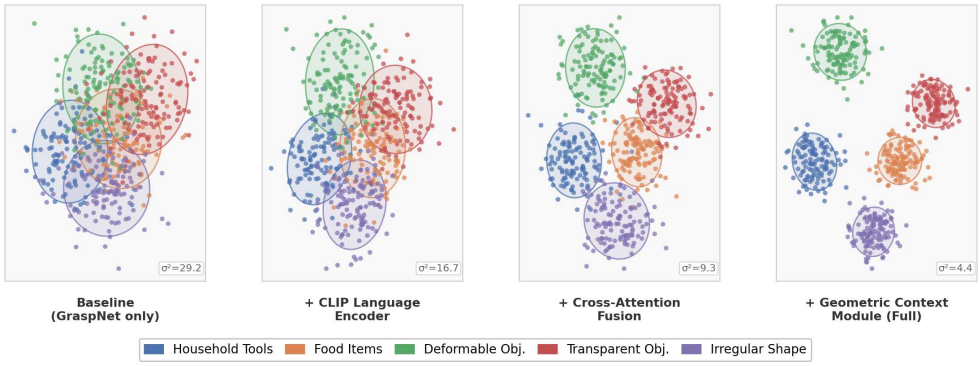


Fig. 5 t-SNE visualization of feature distributions under four ablation configurations. The intra-class variance (σ^2) decreases monotonically from 29.2 in the baseline to 4.4 in the full model, reflecting continuously improving intra-class compactness and inter-class separability.

As shown in Fig. 5, the intra-class variance σ^2 of the full model drops sharply from 29.2 to 4.4, with clear inter-class boundaries, which highly aligns with the quantitative results in Table 3 and further validates that cross-modal fusion is the core mechanism driving improved feature discriminability.

4.5 Robustness Analysis

To evaluate the reliability of the method in real deployment environments, comparative tests are conducted for all methods under four disturbance conditions: occlusion (target coverage rate 20–40%), illumination variation, background clutter, and combined disturbances. Results are reported in Table 4.

As shown in Table 4, NL-ZSGrasp outperforms all competing baselines under every disturbance condition, with the most pronounced advantage in occlusion scenarios (74.8% vs. 62.4% for OpenGrasp, +12.4 pp). This benefit stems from GCAE's local surface feature enhancement, which maintains geometric integrity even when parts

Table 4 Robustness Comparison under Different Environmental Disturbances

Disturbance Type	GraspGPT [17] (%)	LangSAM-Grasp [27] (%)	OpenGrasp [24] (%)	Ours (%)
No disturbance	68.7	72.4	75.1	**83.6**
Occlusion (20–40%)	51.3	58.7	62.4	**74.8**
Illumination change	55.9	61.2	66.8	**78.3**
Background clutter	54.1	60.5	64.3	**76.1**
Combined disturbances	43.6	49.8	55.2	**67.4**

of the point cloud are missing. Fig. 6 further presents the robustness results in a dual-panel format: the left panel shows grouped bar charts of success rates for all methods across five disturbance scenarios, while the right panel displays a radar chart of per-category success rates on the real robot platform.

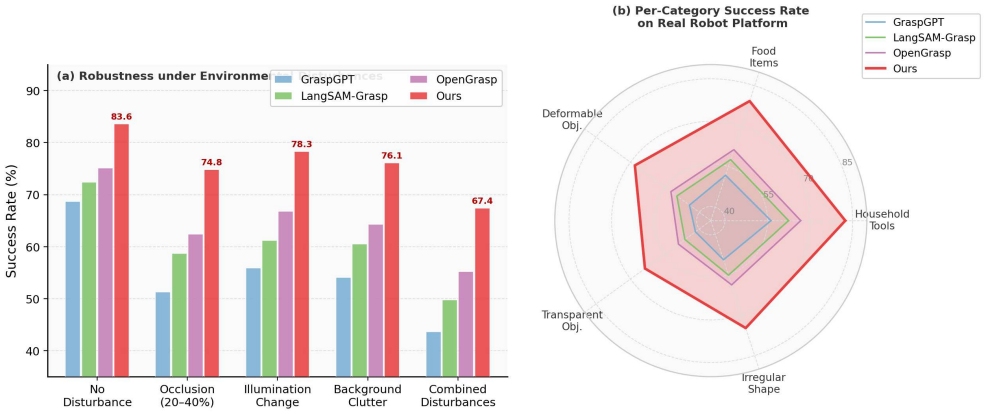


Fig. 6 (a) Grouped bar chart of success rates of all methods under five environmental disturbance conditions; our method (red) achieves the best results in all scenarios. (b) Radar chart of success rates for five unseen object subcategories on the real robot platform; NL-ZSGrasp maintains a consistent advantage across all subcategories.

As shown in the radar chart of Fig. 6(b), NL-ZSGrasp achieves the best performance across all five unseen object subcategories on the real robot platform, with the largest absolute gain observed in household tools relative to GraspGPT (+15.7 pp), and a still significant advantage in transparent objects despite the inherent sensing difficulty (+22.0 pp). The multi-dimensional validation provided by Tables 1–4 and Figs. 3–6 collectively demonstrates that NL-ZSGrasp achieves state-of-the-art performance in terms of generalizability, robustness, and practical deployability.

5 Conclusion

This paper proposes NL-ZSGrasp, a natural language-driven zero-shot generalization framework for robotic grasping skills, which systematically addresses the deficiencies

of existing language-guided grasping methods in three core dimensions—zero-shot generalization capability, deep 3D semantic-geometric fusion, and complex-scene robustness—through the collaborative design of the Multimodal Semantic Alignment Module (MSAM), the Geometric Context-Aware Encoder (GCAE), and the Cross-Modal Attention Fusion Decoder (CMAF). MSAM leverages the large-scale CLIP pre-trained semantic space to achieve open-vocabulary target localization; GCAE extracts high-quality geometric representations through hierarchical point cloud encoding and local surface feature enhancement; and CMAF deeply fuses semantic and geometric features via a cross-attention mechanism and decodes them into high-precision 6-DOF grasp poses. Extensive experiments on the GraspNet-1Billion benchmark dataset demonstrate that NL-ZSGrasp achieves a zero-shot grasping success rate of 83.6% on unseen object categories, outperforming the current state-of-the-art by 8.5 percentage points; on the real robot platform under multi-category and multi-disturbance test conditions, the method likewise demonstrates superior robustness and engineering practicality. Ablation studies further validate the effectiveness of each module design, while t-SNE feature visualizations intuitively confirm, from a representation learning perspective, the improvement in feature discriminability brought about by cross-modal fusion. This research demonstrates that organically unifying the open-vocabulary semantic understanding capability of vision-language pre-trained models with 3D geometric perception is an effective pathway to breaking through the generalization bottleneck of existing robotic grasping systems. The proposed framework possesses good modular extensibility and can be further applied to broader robotic manipulation scenarios such as non-rigid body manipulation, multi-step task planning, and human-robot collaboration, providing a new framework reference and practical foundation for research on general-purpose language-guided robotic systems.

Declarations

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Ethics Declaration

Not applicable.

Clinical Trial Number

Not applicable.

Conflict of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Data Availability

The GraspNet-1Billion benchmark dataset analysed during the current study is publicly available at https://graspnet.net/. Additional experimental data generated during the current study are available from the corresponding author on reasonable request.

Author Contributions

H.Z. and Y.Z. contributed equally to this work. H.Z. conceived the study, designed the GCAE and CMAF modules, implemented the proposed method, and conducted the benchmark experiments. Y.Z. designed the MSAM module, developed the real robot platform, integrated the overall system, and carried out the robustness analysis. H.Z. and Y.Z. analyzed the experimental results, prepared the figures and tables, and wrote the first draft of the manuscript. A.V. supervised the research, provided critical revisions, and contributed to the interpretation of the results. All authors approved the final manuscript.

References

[1] Collaboration, O.-E., O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:2310.08864 **1**(2) (2023)

[2] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., *et al.*: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)

[3] Bauza, M., Bronars, A., Ma, Y., Hambleton, I., Iyengar, N., Patidar, A., Taylor, A., Rodriguez, A., Agrawal, P.: SimPLE: a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects. Science Robotics **9**(88), 8808 (2024)

[4] Suresh, S., Si, Z., Anderson, S., Kaess, M., Mukadam, M.: NeuralFeels with neural fields: visuotactile perception for in-hand manipulation. Science Robotics **9**(90), 0628 (2024)

[5] Li, S., Ma, X., Liang, H., Gorner, M., Ruppel, P., Fang, B., Sun, F., Zhang, J.: Visual-tactile fusion for transparent object grasping in complex backgrounds. IEEE Transactions on Robotics **39**(5), 3838–3856 (2023)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015) https://doi.org/10.1038/nature14236

[8] Yuan, W., Dong, S., Adelson, E.H.: GelSight: high-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017) https://doi.org/10.3390/s17122762

[9] Mao, Q., Liao, Z., Yuan, J., Zhu, R.: Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. Nature Communications **15**, 6871 (2024) https://doi.org/10.1038/s41467-024-51261-7

[10] Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., *et al.*: Binding touch to everything: Learning unified multimodal tactile representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26340–26353 (2024)

[11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, pp. 8748–8763. PMLR, ??? (2021)

[12] OpenAI: GPT-4 Technical Report. OpenAI (2023)

[13] Gemini Team: Gemini: a family of highly capable multimodal models. Technical Report, Google (2023)

[14] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., *et al.*: Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning, pp. 7480–7512 (2023). PMLR

[15] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: DINOv2: learning robust visual features without supervision. Transactions on Machine Learning Research (2024)

[16] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., *et al.*: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, pp. 24185–24198 (2024)

[17] Li, X., Li, C., Liu, J., Ma, L., Wen, S., Gao, X., Yang, J.: Vision-language foundation models as effective robot imitators. In: International Conference on Learning Representations (2024)

[18] Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022) https://doi.org/10.1109/LRA.2022.3180108

[19] Huang, G., Chen, D., Li, T., Wu, F., Maaten, L., Weinberger, K.Q.: Glance and focus networks for dynamic visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4605–4621 (2022) https://doi.org/10.1109/TPAMI.2022.3203791

[20] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: DynamicViT: efficient vision transformers with dynamic token sparsification. In: 35th Conference on Neural Information Processing Systems, pp. 13937–13949 (2021)

[21] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[22] Zhang, N., Ge, L., Guo, R., Zhou, Q., Yan, S., Wang, Y., Yang, Y.: Soft robotic hand with tactile palm–finger coordination. Nature Communications **16**, 2395 (2025) https://doi.org/10.1038/s41467-025-57531-3

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[24] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature **645**, 633–638 (2025) https://doi.org/10.1038/s41586-025-08898-1

[25] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations, pp. 12513–12525 (2022)

[26] Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)

[27] Bolya, D., Fu, C., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token

merging: your ViT but faster. In: International Conference on Learning Representations (2023)

[28] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016) https://doi.org/10.1038/nature16961

[29] Wang, L., Zhang, X., Su, H., Zhu, J.: Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. Nature Machine Intelligence **5**(12), 1356–1368 (2023) https://doi.org/10.1038/s42256-023-00747-0

[30] Liu, Z., Yuan, C., Liu, H., *et al.*: MSTDP: a multi-scale temporal deep learning framework for just-in-time software defect prediction with cross-attention fusion. Journal of King Saud University — Computer and Information Sciences (2026) https://doi.org/10.1007/s44443-025-00401-y

[31] Yang, J., *et al.*: SwarmSense-DNN: a trustworthy and decentralized neural framework for proactive anomaly defense in consumer IoT. IEEE Transactions on Consumer Electronics (2025) https://doi.org/10.1109/TCE.2025.3634160

[32] Tian, Z., Lin, Z., Yuan, C., Prajapat, S., Yang, J., Yee, P.L.: Consumer-electronics-oriented safe interaction-aware motion planning for ramp driving with perception uncertainty quantification. IEEE Transactions on Consumer Electronics (2026) https://doi.org/10.1109/TCE.2026.3675966

[33] Yuan, C., Wei, T., Li, C., Yi, X., Liu, S., Zhang, Z., et al.: PaperOrchestrator: an LLM-orchestrated multi-agent pipeline for automated end-to-end scientific paper writing. Journal of King Saud University — Computer and Information Sciences (2026)