Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077916

ARTICLE

Secondary Realignment: An Embodied Intelligent Operational Framework Integrating Vision-Language and Action Two-Stage Models

Jinjiang Lin, Yuan Lu, Han Li, Xiaolong Cai, Enyi Chen and Jiansheng Guan*

School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen, China

*Corresponding Author: Jiansheng Guan. Email: jsguan@xmut.edu.cn

Received: 19 December 2025; Accepted: 03 March 2026; Published: 08 May 2026

ABSTRACT: Manipulating objects based on verbal commands in cluttered environments remains a critical challenge in robotic arm research. Verbal commands possess high semantic abstraction, while precise grasping and placement actions rely on fine-grained geometric perception. The disparity between these two domains is the primary cause of operational errors. Particularly in certain cluttered scenarios, visual-spatial noise and background redundancy further disrupt attention distribution, significantly degrading the generalization capabilities of existing methods in unseen environments. To address these issues, this paper proposes the Secondary Realignment (SR) framework. It decouples vision-language alignment and vision-action alignment into two stages, mitigating semantic-geometric discrepancies through a hierarchical approach to substantially reduce errors in cross-modal mapping. Simultaneously, to address noise and redundancy in visual-language features, we design a Deep Sparse Self-Attention (DSSA) module. This module dynamically fuses sparse and dense attention mechanisms through self-learning parameters, adaptively enhancing relevant features while suppressing irrelevant noise. Extensive simulation experimental results demonstrate that compared to the state-of-the-art method A2, our approach achieves 9.7%, 9.9%, and 17.6% higher task success rates in grasping, placing, and pick-and-place tasks, respectively, further validating its effectiveness.

KEYWORDS: Robot grasping; visual language model; language-conditional grasping; attention mechanism

1 Introduction

Performing language-conditioned grasping and placement tasks in cluttered environments represents one of the most challenging capabilities in current embodied intelligence research. Robots must infer task intent from natural language, identify target objects within complex 3D scenes, understand their geometric structures, and generate high-precision 6-DoF actions. However, this cross-domain "vision-language-action" chain suffers from significant semantic-geometric inconsistencies: language provides abstract semantics, while action generation relies on fine-grained spatial relationships. Once misalignment occurs between them, errors propagate and amplify through subsequent modules.

Existing approaches include end-to-end policy networks and modular vision-language operating systems. End-to-end approaches [1–3] perform well in simple scenes but are prone to overfitting and lack robustness when encountering unseen objects or novel layouts. Modular methods typically employ visual-language models and related large-model frameworks [4–6] for instruction interpretation, semantic grounding [7,8], and then integrate them with separate action generation or grasp planning modules [9]. However, modular systems struggle with cascading errors when combining several foundation models in a zero-shot setting and their single-stage cross-modal alignment often leads to semantic drift. For instance,

correct semantic alignment but misaligned action placement is a common failure mode. Furthermore, in highly cluttered scenes, dense point clouds contain numerous occluded points and noisy regions, where standard attention mechanisms are often disrupted by irrelevant structures, further degrading alignment quality and action stability.

To address the aforementioned challenges, we propose a Secondary Realignment (SR) framework. Its core concept transforms cross-modal alignment from a "single-stage strong coupling" to a "phased weak coupling." In the first stage, the model constructs coarse-grained semantic correspondences between linguistic and visual features, ensuring reliable identification of task-relevant regions. In the second stage, action-based priors are introduced to finely map semantically enhanced visual representations to the 6-DoF action space, thereby mitigating the direct interference of semantic misalignment on action prediction. This hierarchical alignment mechanism enables the model to robustly handle semantic variations and geometric changes in unseen scenes. Concurrently, the Deep Sparse Self-Attention (DSSA) module is introduced. It adaptively suppresses task-irrelevant regions in visual-language features by learnably fusing sparse and dense attention, significantly enhancing robustness in scenarios with severe occlusions or complex backgrounds.

Based on the aforementioned problem background and methodological rationale, the main contributions of this paper are summarized as follows:

(1) We propose the Secondary Re-Alignment (SR) framework, which decouples language–vision and vision–action cross-modal alignment into two complementary stages. This mechanism alleviates cascading errors caused by semantic–geometric inconsistencies and significantly enhances zero-shot generalization capabilities in unseen scenarios.
(2) Introduces the Deep Sparse Self-Attention (DSSA) module, which fuses sparse and dense attention to adaptively filter semantic noise in cluttered visual-language features, enabling the model to maintain stable key feature focus in complex scenes.
(3) Demonstrated substantial performance gains across large-scale experiments in Grasp, Place, and Pick-Place tasks and diverse scenarios. The SR framework achieved improvements of 9.7%, 9.9%, and 17.6%, respectively, in unseen scenarios, validating the proposed method's significant advantages in cross-modal alignment quality, action reliability, and language-conditional embodied operation performance.

2 Related Work

2.1 Task-Oriented Operations

Unlike simple grasping, task-oriented manipulation requires robots to understand "what to grasp" and "how to use it". Traditional approaches typically rely on large-scale annotated datasets, achieved through part segmentation [10], probabilistic modeling [11], or availability detection [12]. To overcome data limitations, recent research has shifted toward leveraging the zero-shot capabilities of pre-trained large models. While these models endow robots with powerful semantic understanding, existing "semantic-geometric" mappings often remain loose. Most approaches stop at object recognition, lacking effective alignment mechanisms when translating semantic intent into precise 3D motion planning. This leads to models correctly understanding semantics but failing in physical execution. Furthermore, existing geometric decomposition approaches like ShapeGrasp [13], while offering novel insights, face application limitations. Addressing this "semantic-action" gap, this paper establishes a hierarchical mapping from language to point cloud to action, effectively bridging the shortcomings of existing methods in complex task execution.

2.2 Grasping in Cluttered Environments

Achieving robust grasping in cluttered, unstructured environments is a fundamental challenge in embodied intelligence. Early traditional methods relied on geometric analysis and assumptions of structured environments, but their generalization capabilities were severely limited when encountering dynamic scenarios such as stacked objects and severe occlusions [14]. With the advancement of deep learning, methods based on convolutional neural networks (CNNs) enhanced scene adaptability to some extent by extracting deep visual features [15–17]. Simultaneously, the introduction of reinforcement learning (RL) enabled robots to optimize grasping strategies through trial-and-error mechanisms, enhancing dynamic interaction capabilities. However, these approaches universally suffer from data inefficiency and generalization bottlenecks: they not only depend on large-scale labeled datasets but also exhibit steep performance degradation when encountering unseen objects [18,19]. Although recent works like Sim-Grasp [20] have attempted to improve performance through enhanced network architectures, the challenge remains: how to effectively handle noise interference in cluttered environments and maintain high robustness without massive data support. To address this, this paper enhances the model's ability to capture key features by adjusting attention distribution, thereby reducing dependence on data volume.

2.3 Cross-Modal Alignment

Cross-modal alignment aims to map linguistic semantics to visual features and translate them into executable 6-DoF actions, forming the core of language-conditioned embodied operations. Existing research predominantly employs Transformers to jointly encode linguistic and scene features for end-to-end prediction, or leverages large-scale embodied models [21,22] to enhance semantic understanding in complex tasks. However, these approaches generally adopt a single-stage, tightly coupled strategy, modeling multimodal features within the same attention space. While this tight coupling unifies feature representations, it also leads to error propagation: minor deviations in vision-language associations directly propagate to the action space, causing semantic drift and execution misalignment. Although recent efforts have attempted to enhance robustness through Large Language Model (LLM) assistance, hierarchical task decomposition, or constructing continuous semantic fields (e.g., LERF-TOGO [23]), most remain limited to improving semantic understanding and lack precise alignment mechanisms from the semantic space to the geometric action space. Particularly in complex point cloud environments, semantic noise continues to significantly disrupt the stability of action planning.

2.4 Robust Perception for Manipulation

In the field of robotic manipulation, particularly in handling complex and cluttered environments, perception robustness poses a critical challenge. The document discusses several methods and technologies aimed at enhancing perception accuracy, yet each approach suffers from distinct limitations. 2D-based approaches [24,25] leverage large-scale visual pretraining to extract rich semantics; however, they inherently lack explicit 3D spatial understanding, which hinders their effectiveness in tasks requiring precise geometric interaction. Conversely, point cloud-based methods [26,27] capture explicit 3D structures but are frequently hampered by inherent sparsity and sensor noise, making it difficult to maintain semantic consistency across diverse views. Voxel-based representations [28] reduce sparsity by discretizing space for structured reasoning, yet they incur high computational costs. Multi-view RGB-D approaches [23,29] integrate dense 2D semantics with geometry via early fusion or auxiliary supervision, yet such shared feature spaces remain vulnerable to sensor noise and often lack precise spatial correspondence for fine-grained interaction. To address these limitations, we decouple semantics and geometry and construct geometric representations by fusing complementary priors and raw depth together with low-level spatial cues for precise manipulation.

3 Method

3.1 Model Architecture

The overall architecture of the SR framework is illustrated in Fig. 1. To achieve effective integration among language, vision, and action, the system is divided into three core modules: preprocessing, two-stage cross-modal alignment, and action planning. This design aims to progressively transition from semantic understanding to executable actions, thereby enhancing the robot's task robustness in complex scenarios.

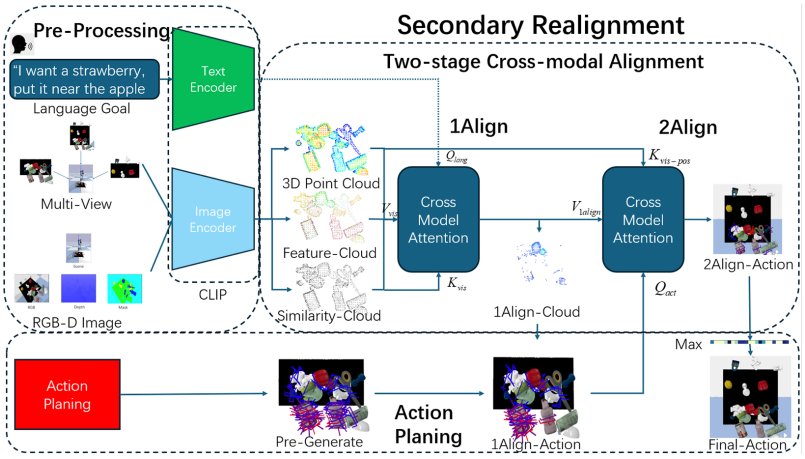

Figure 1: System overview.

We leverage the zero-shot generalization capability of foundation models to construct 3D visual representations that convey semantic and task-relevant information. During the preprocessing stage, the system receives natural language instructions (e.g., "I want a strawberry, put it near the apple") along with multi-view RGB-D images. Text input is converted into linguistic feature vectors via a text encoder; visual input is parsed by a CLIP-based visual encoder [30] into visual feature maps and semantic similarity maps. Concurrently, the action prior module generates a set of initial action candidates, providing an action hypothesis space for subsequent fine-grained alignment.

Subsequently, SR performs cross-modal alignment at two levels. In the first stage, linguistic features serve as queries Q, while visual features act as key K and value V. Through cross-modal attention, it generates a language-driven semantic point cloud representation (1Align-Cloud), achieving coarse-grained "vision-language" association. The second stage further incorporates action candidates into the attention mechanism. Using candidate actions as queries and visual-position features as keys, it performs fine-grained alignment between actions and spatial semantics, yielding 2Align-Action. This stage strengthens the correspondence between semantic labels and 3D positions, thereby providing more precise structural information for action selection.

After obtaining the alignment results from the two stages, the motion planning module GRASPNET [31] infers multiple sets of executable grasping/placement actions 1Align-Action and 2Align-Action, and selects the optimal action through semantic consistency scoring. This closed-loop process ensures that the final generated 6-DoF actions satisfy both the semantic constraints of the linguistic commands and the geometric structure of the scene, significantly enhancing the robot's task execution capability in cluttered environments.

3.2 Preprocessing Module

3.2.1 Visual Language Encoder

To achieve efficient and consistent visual-linguistic feature integration, this paper employs the Contrastive Language-Image Pre-training (CLIP) framework to generate cross-modal shared semantic representations. In terms of visual feature construction, as shown in Fig. 2, a multi-scale 3D feature generation strategy is designed. Specifically, the ViT-L/14 encoder from CLIP extracts dense 2D patch-level features from multi-view RGB-D images. These 2D pixel features are then elevated into 3D space via a camera projection model, forming a 3D feature point cloud with true geometric structure. This process not only preserves semantic information from the images but also explicitly encodes spatial relationships within the scene, providing a precise geometric foundation for subsequent cross-modal alignment.

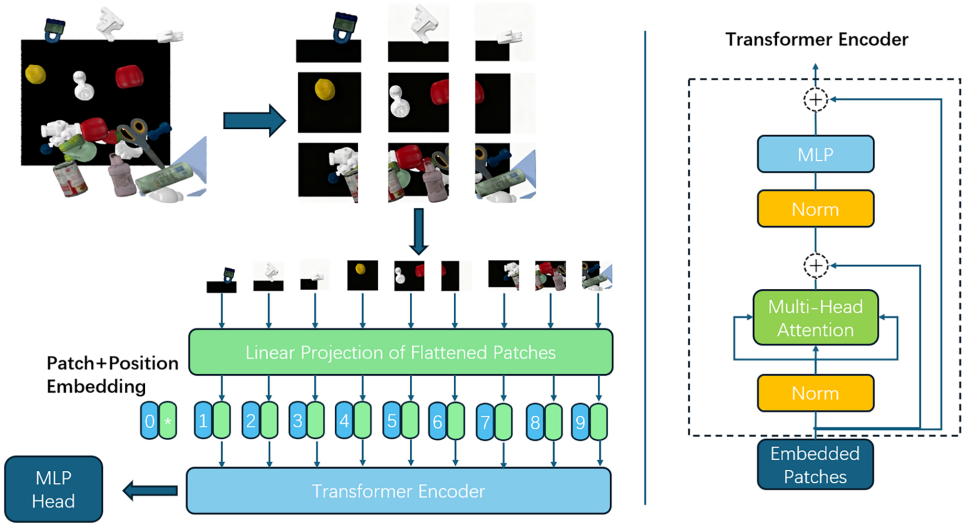

Figure 2: Visual coding.

For language processing, the CLIP text encoder shares a semantic space with the visual encoder, mapping natural language instructions to the same embedding space. This achieves a natural alignment between instruction semantics and scene representations. Both text and images are divided into fixed-size patches. After linear embedding and positional encoding, these patches are fed into a standard Transformer encoder to model their global semantic relationships. This design establishes a robust correspondence between visual and linguistic features within a unified semantic space, laying the foundational structure for the subsequent two-stage cross-modal alignment mechanism.

3.2.2 Spatial-Level Freedom in 3D Spatial Perception

To construct a consistent 3D semantic representation across multiple viewpoints, this study maps 2D features from multi-lens RGB-D images into a unified 3D world coordinate system. Specifically, for each candidate 3D point, we first project it onto the corresponding pixel location in each view using camera intrinsic and extrinsic parameters. We then compare the deviation between its projected depth and the measured depth map value. This deviation is used to construct an exponentially decaying confidence weight, which measures the reliability of observation for that point from a specific viewpoint.

At high-confidence viewpoints, sub-pixel features are sampled from 2D feature maps via bilinear interpolation to obtain more precise local descriptions. Subsequently, sampled features from all viewpoints

are weighted and fused according to their confidence levels, thereby integrating complementary information from multiple perspectives. Ultimately, each 3D point is associated with a fused multimodal feature vector, endowing it with both semantic consistency and geometric alignment. This yields a structurally stable and visually complete three-dimensional semantic feature cloud. The corresponding process can be formally represented as follows:

$$\omega_{depth}^{(k)}(i) = \exp\left(\frac{\max\left(0, \left|d_{meas}^{(k)}(i) - d_{true}^{(k)}(i)\right|\right)}{\mu}\right) \tag{1}$$

$$f_{\text{interp}}^{(k)}(i) = BilinearInterpolate\left(F^{(k)}, P_{uv}^{(k)}(i)\right) \tag{2}$$

$$f_{final}(i) = \frac{\sum_{\text{k=1}}^{K} v^{(k)}(i) \cdot \omega_{depth}^{(k)}(i) \cdot f_{\text{interp}}^{(k)}(i)}{\sum_{\text{k=1}}^{K} v^{(k)}(i) \cdot \omega_{depth}^{(k)}(i)} \tag{3}$$

3.3 Two-Stage Cross-Modal Alignment

3.3.1 Cross-Modal Attention Mechanism

To adapt action generation to linguistic and visual conditions, we further modulate the unconditional action prior using task-relevant visual-linguistic priors. To this end, we employ the Transformer attention mechanism, as shown in Fig. 3:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

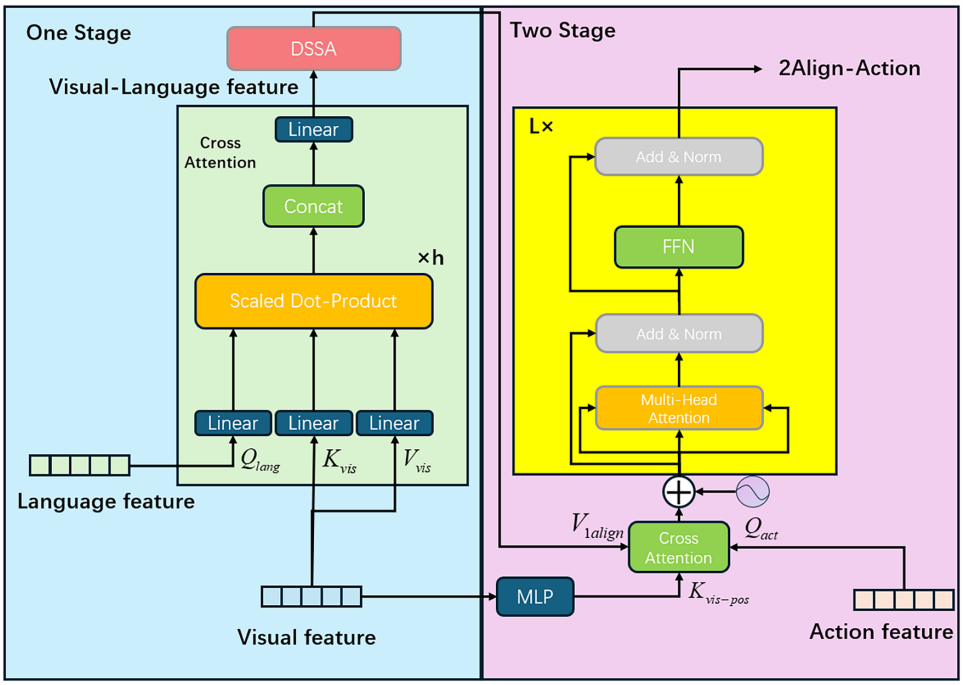

Figure 3: Two-stage cross-modal alignment.

Its core idea is to decompose the cross-modal alignment process into two stages: language-driven semantic alignment (1-Align) and action-driven geometric alignment (2-Align). This approach mitigates the structural differences between linguistic semantics and 3D geometry, reducing semantic drift and geometric misalignment.

Coarse Semantic Localization. The language-driven semantic alignment stage aims to achieve a coarse-grained semantic localization task-relevant regions in the scene by filtering the visual features according to linguistic intent. Specifically, it uses linguistic features as queries (Q_{lang}), while visual features serve as keys (K_{vis}) and values (V_{vis}). This cross-modal attention employs a multi-head attention mechanism that computes the scaled dot-product, concatenates the results, and projects them through a linear layer, ultimately outputting the semantically enhanced representation. To further enhance cross-modal fusion, the DSSA module is introduced after this first stage. Through attention weight sparsification and structural guidance, it significantly strengthens the model's ability to focus on key visual-linguistic regions. This representation functions as a semantic mask, retaining only task-relevant regions and effectively filtering out irrelevant information, thereby providing attention guidance for subsequent geometric alignment.

$$V_{1align} = \text{softmax}\left(\frac{Q_{\text{lang}}K_{vis}^{T}}{\sqrt{d}}\right)V_{vis} \tag{5}$$

Fine Geometric Adaptation. Building upon semantic alignment, the action-driven geometric alignment phase further achieves fine-grained geometric adaptation from visual to action. The structured action priors are evaluated against semantically constrained local geometry to assess action executability. The original visual features are mapped through a Multilayer Perceptron (MLP) to generate visual-position features, which serve as the keys $K_{\text{vis-pos}}$. Using the visual-language features output from 1-Align as the value $V_{1allign}$ and action priors as the query Q_{act}, it performs geometric alignment on candidate 6-DoF poses one by one through cross-attention

$$S_{act} = \text{softmax}\left(\frac{Q_{act}K_{\text{vis-pos}}^{T}}{\sqrt{d}}\right)V_{1align} \tag{6}$$

To further refine these cross-modal representations, the output of the cross-attention module is combined with positional encodings and fed into a stacked L-layer Transformer encoder. Each encoder layer consists of Multi-Head Self-Attention and Feed-Forward Networks with residual connections and layer normalization.

For attention-based alignment, action priors are represented as compact 7-dimensional pose vectors and embedded via a three-layer MLP into a unified latent space of dimension $D = 512$, forming a query tensor of shape that interact with semantically filtered scene features through cross-attention.

Since action queries align only within highly relevant regions constrained by semantic masks, rather than the entire point cloud. This approach effectively mitigates the amplified impact of semantic misguidance on geometric planning. In our framework, each candidate action is evaluated independently but under identical semantic and geometric constraints, so that the resulting action-geometric scores are directly comparable.

3.3.2 DSSA Module

To effectively suppress noisy features while preserving global dependency modeling capabilities, this paper proposes a Deep Sparse Self-Attention (DSSA) mechanism. Traditional Transformers perform fully

connected operations on all tokens during attention computation. While this captures long-range relationships, it often introduces substantial interference from irrelevant regions, leading to redundant computations and noise accumulation. To mitigate this issue, DSSA employs a dual-branch architecture combining dense and sparse attention to balance information retention and noise filtering. Given the visual-language feature $X \in R^{N\times C}$, from the first-stage alignment, it is linearly projected to obtain the queries Q, keys K, and values V:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{7}$$

Among these, W_Q, W_K, and $W_V \in R^{C\times \mathrm{d}}$ are projection matrices. The attention calculation can be defined as:

$$A = \mathrm{f}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{8}$$

B denotes the learnable relative positional offset, and $f(x)$ is the scoring function. It is worth noting that subsequently, weight calculations are performed in parallel for different heads, which are then concatenated and fused through linear projection.

Dense Convolutional Self-Attention (DCSA) Branch: This branch follows a traditional design, applying convolutions to the conventional attention score $\frac{QK^T}{\sqrt{d}}$. This enables the model to learn local spatial patterns and enhances its ability to focus on relevant regions.

$$DCSA = SoftMax\left(\mathrm{Conv}\left(\frac{QK^T}{\sqrt{d}}\right) + B\right) \tag{9}$$

This branch can fully preserve global information while remaining equally sensitive to noisy regions.

Sparse Self-Attention (SSA) Branch [32]: To mitigate the influence of irrelevant regions, a sparse attention mechanism based on squared ReLU is introduced. This mechanism masks negative similarity scores while amplifying high-similarity connections, thereby yielding sparser and more robust attention patterns. This design fits visual-language characteristics that informative regions are sparse and localized, while most points are background clutter or noise. Compared to softmax-based attention, squared ReLU retains strong responses without global normalization, preserving sharp geometric boundaries and avoiding irrelevant point amplification.

$$SSA = \mathrm{Re}LU^2\left(\frac{QK^T}{\sqrt{d}} + B\right) \tag{10}$$

Although SSA can effectively filter noise, its sparsity may lead to information loss, resulting in feature representations that are insufficient to support subsequent tasks. Relying solely on either approach has limitations: DCSA introduces noise, while SSA loses information. Therefore, we propose an adaptive fusion strategy to dynamically balance the two.

$$DSSA = (\alpha_1 * SSA + \alpha_2 * DCSA)\, V \tag{11}$$

Here, $\alpha_1, \alpha_2 \in R^1$ are layer-wise learnable parameters updated via standard backpropagation. They are normalized via a Softmax function to ensure numerical stability during training.

$$a_{\mathrm{n}} = \frac{e^{\beta_n}}{\sum_{i=1}^{N} e^{\beta_i}}, n = \{1, 2\} \tag{12}$$

β_1, β_2 are a trainable parameter. This layer-wise granularity enables the network to adjust feature sparsity according to task requirements: favoring sparse attention in noisy scenes while allowing denser information flow when preserving details is essential, thereby achieving a dynamic balance between noise suppression and information integrity (as shown in Fig. 4).

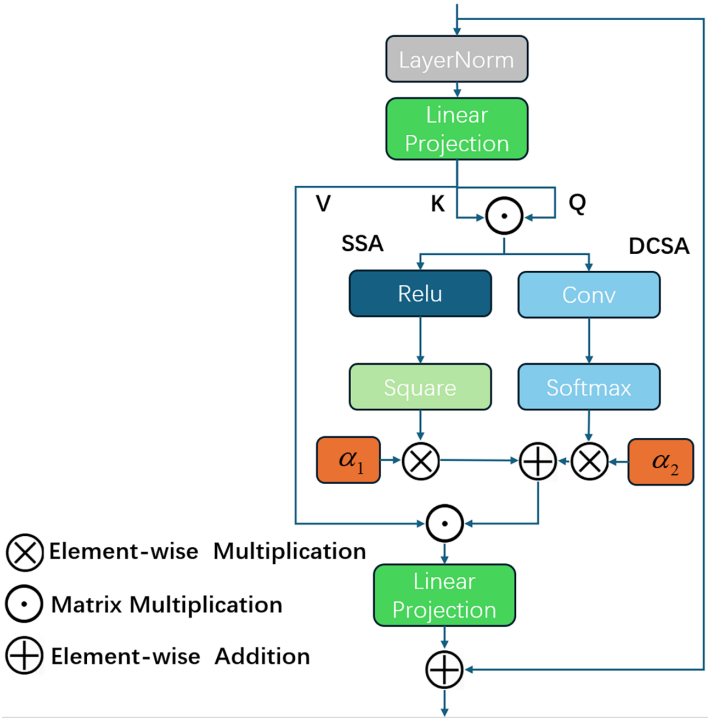

Figure 4: DSSA module.

3.4 Action Planning Module

3.4.1 Action-Guided Pre-Grasping Model

In complex embodied manipulation tasks, planning precise actions directly from high-dimensional multimodal data poses significant challenges. To address this issue, this module aims to provide action priors that drastically reduce the action search space, thereby enhancing the model's planning efficiency and robustness in unstructured environments. To generate a pre-grasping pose $\mathcal{I} = \{I_{\mathrm{i}}\}_{\mathrm{i}=0,1,\ldots,M}$ from an RGB-D image and a language instruction $\mathcal{L}$, we first employ an action foundation model GraspNet [31] to generate a set of L candidate reference actions $A_L(\mathcal{I}) = \{a_k\}_{k=0,1,\ldots,L}$ based on the image. Each candidate action is represented as: $\mathrm{a}_k = [\mathrm{a}_{pos}, \mathrm{a}_{rot}, \mathrm{a}_{open}]$, where $\mathrm{a}_{pos} = (x, y, z)$ is 3D translation, $\mathrm{a}_{rot} = (q_x, q_y, q_z, q_w)$ denotes the unit quaternion and a_{open} indicates whether the end-effector is closed. Based on this, a probabilistic

policy π was constructed to determine the final pre-grasping pose based on these reference actions.

$$\pi\left(a|\mathcal{I},\mathcal{L}\right)=\sum_{k=1}^{L}w\left(a_k|\mathcal{I},\mathcal{L}\right)\delta\left(a-a_k\right) \tag{13}$$

Among these, the weights satisfy the normalization constraint:

$$\sum_{k=1}^{L}w\left(a_k|\mathcal{I},\mathcal{L}\right)=1 \tag{14}$$

where $w\left(a_k|\mathcal{I},\mathcal{L}\right)$ proves the probability of selecting the Kth reference action a_k under visual and linguistic information. We conduct independent action evaluation, so even in complex scenarios with highly overlapping multiple targets, the final policy can still stably and consistently select the action that best matches the semantic intent of the language instruction and has the highest geometric feasibility. Importantly, the role of this module is not to introduce a novel action prior model, but to provide structured, physically plausible candidate actions that can be flexibly replaced by alternative generators. These action priors are subsequently refined and evaluated through the proposed staged semantic–geometric alignment mechanism, ensuring that final action selection is guided by both linguistic intent and local geometric feasibility.

3.4.2 Action Decoding

After obtaining the probability distribution, the model scores the grasping points through the inner product of parameter queries to select one as the next optimal end-effector position a_{pos}. Then, from the previous iteration parameters of the three-layer multi-layer perceptron (MLP), it queries the orientation a_{pos} and open a_{open} of the end-effector, as well as whether the motion planner needs to avoid collisions to achieve pose a_{col}. Finally, the process outputs a precise grasping pose that integrates semantic intent with geometric constraints. The action execution flow is illustrated in Fig. 5.

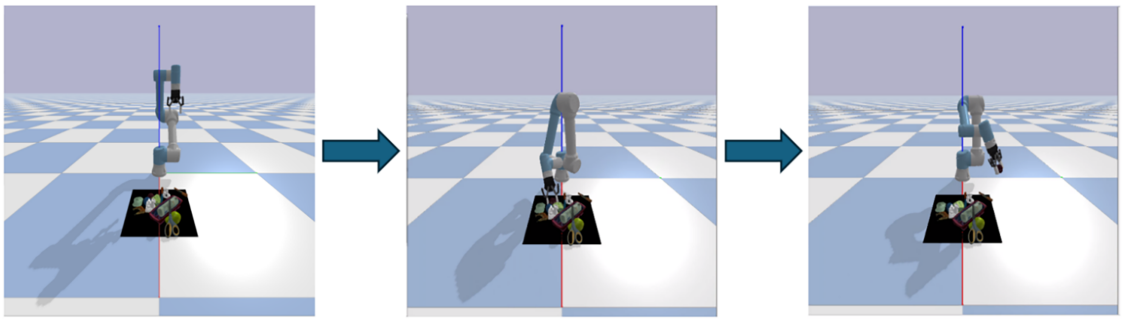

Figure 5: Action execution.

3.4.3 Loss Function

The objective of the task is to maximize the probability of action success for each candidate action $A_L\left(\mathcal{I}_d\right)$ in each demonstration $D=\{\mathcal{I}_{\mathrm{d}},\mathcal{L}_d,a_d\}$. For each step, the expert-demonstrated action is matched to the closest candidate and treated as the positive sample. All remaining physically feasible candidates within the same step are implicitly treated as hard negative samples. The model is trained using a cross-entropy loss over candidate scores, computed per decision step rather than per trajectory.

$$\mathcal{L}_{CE}=-\log\omega\left(a_d|\mathcal{I}_d,\mathcal{L}_d\right) \tag{15}$$

Multimodal ambiguity in language instructions is addressed by supervising actions at the region and relation level rather than enforcing a single deterministic target. In particular, placement demonstrations are generated from valid spatial regions (e.g., *on*, *around*), allowing multiple semantically correct actions to exist.

4 Experiment

In this section, the study conducted a series of experiments to evaluate the system. The objectives of the experiments were: (1) to demonstrate the strategy's effectiveness for language-guided grasping tasks in cluttered environments; (2) to assess the strategy's generalization performance for unseen objects and verbal instructions; (3) to validate the effectiveness of each component module.

4.1 Experimental Setup

Simulated Environment. Demonstration data was collected using the model-based expert planner for the UR5 arm in PyBullet [33]. Three statically mounted cameras (M = 3) overlook the tabletop: one facing front at 90°, one viewing diagonally downward at 45°, and one viewing diagonally downward at 45°, designated as the top, left, and right cameras, respectively. For each camera, we simulate a depth sensor using the intrinsic parameters and noise characteristics of an Intel RealSense L515. The guidance model originates from GraspNet [31].

Data Collection. To train the Pick and Place tasks, this paper constructed a dataset comprising 5000 demonstration, from which approximately 6500 successful operation samples were selected—3400 for picking and 3100 for placing. Each demonstration provides a supervision signal for a single decision step and serves as a positive action sample during imitation learning. During data collection, distinct scene configuration strategies were designed for each task: For the pick task, 15 objects were randomly stacked within the workspace to simulate cluttered environments. A model-based expert policy selected the closest grasping point to the target. For the place task, to ensure sufficient placement space, 8 objects were arranged with a minimum center-to-center spacing of 0.1 m. The model-based expert policy first identified a valid placement region based on reference objects and spatial relationships, then randomly sampled the final placement point within this region, yielding positive placement actions that satisfy both semantic and geometric.

Training Setup. This study employs an imitation learning framework based on ViT-L/14. Its text encoder adopts the Transformer architecture with the following configuration: width 768, 8 attention heads, and 1 layer. The action decoder consists of a 3-layer perceptron. During training, both the pre-trained weights of ViT-L/14 and the parameters of the action model remain fixed. The network was trained by minimizing cross-entropy loss over 200 epochs. All experiments were conducted on a single NVIDIA GTX 3080ti GPU using the ADAM optimizer, with an initial learning rate of 1e−4 and a batch size of 32.

Test Setup. To comprehensively evaluate model performance, a simulated test comprising three task categories was designed: grasping, placing, and pick-and-place. Each task category includes both seen-object and unseen-object test scenarios to assess the model's generalization capability and adaptability to novel instances, respectively. Here, "seen objects" refer to instances encountered in the training dataset, while "unseen objects" denote novel instances not encountered during training. For the placement task, some unseen objects appear in pairs with unseen relationships to test more complex generalization capabilities (as shown in Figs. 6 and 7).

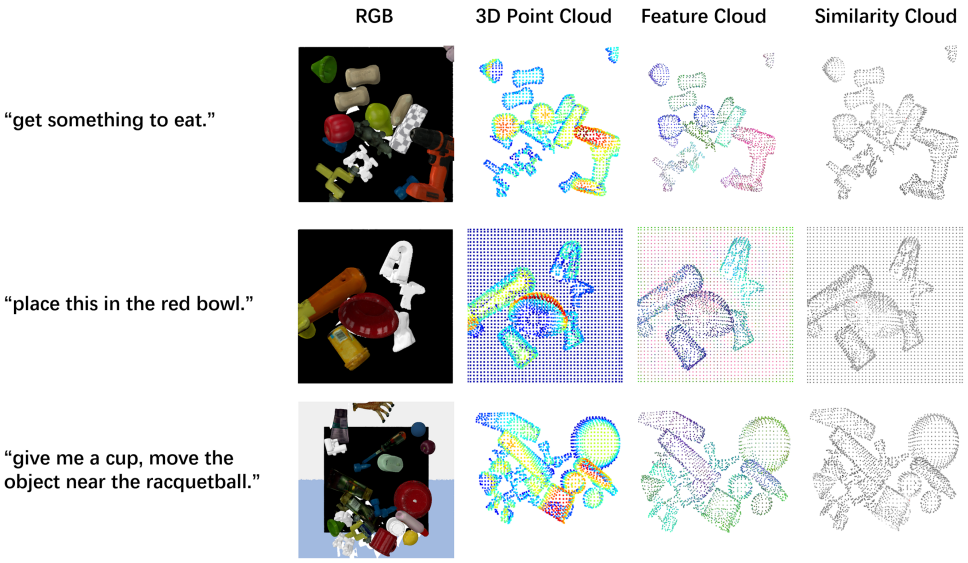

Figure 6: Seen test scenario.

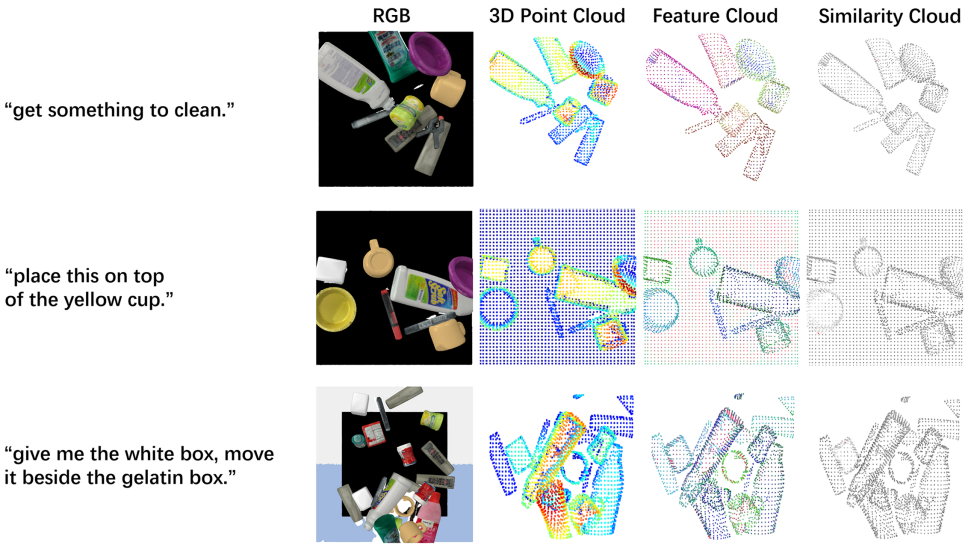

Figure 7: Unseen test scenario.

The experimental design in this paper aims to comprehensively evaluate the generalization and robustness of the policy across tasks of varying complexity. All experiments in this work are conducted in simulation. For the grasping task, a highly challenging scenario was constructed: 15 objects randomly stacked to form a high-density cluttered environment, forcing the policy to learn to reach targets by displacing obstacles. Its generalization capability was validated by testing on 10 seen and 5 entirely novel object arrangements. For the placement task, the study focused on the decision stability of the policy. The scene comprised 8 objects to ensure sufficient operational space. Robustness was tested not only across 20 seen and 10 unseen layouts but also by introducing random perturbations to actions, simulating real-world execution errors to assess the reliability of single-step decisions. The final pick-and-place composite task integrated

these challenges, requiring the policy to perform continuous operations across distinct grasping (left) and placement (right) zones in scenarios with 8 seen and 4 unseen objects.

Success Criteria. In simulated environments, environmental feedback is directly utilized to determine whether a grasping action is successful. If the similarity between the visual segment of the grasped object and the verbal command exceeds a preset threshold, the target object is deemed successfully grasped.

Evaluation Metrics. To quantify model performance, the study established two primary evaluation metrics, with each test case run 15 times to compute averages. The first is task success rate, defined as the average percentage of successfully completed tasks. For grasping tasks, success is defined as grasping the target within 8 attempts; for placement tasks, success is defined as placing the object in the correct region on a single attempt; and for pick-and-place composite tasks, both phases must be successful. Second is the planning steps metric, representing the average number of actions required to complete a task. This metric measures efficiency only for grasping and pick-and-place tasks, not for single-step placement tasks.

4.2 Comparative Experiments

To ensure a rigorous and fair evaluation, we benchmarked SR against other methods under identical sensor configurations and simulated environments. Crucially, for learnable baselines with similar architectures, specifically CLIPort [34] and A^2 [35], we explicitly fine-tuned them on our dataset to adapt to the task.

Grasping task. The CLIPort [34] is an end-to-end approach that leverages CLIP to construct semantic concepts from raw images and directly learns actions by predicting pixel-level grasp heatmaps. CLIP-Grounding employs a two-stage strategy: first, using CLIP to compute similarity between language instructions and all object bounding boxes for target localization, then randomly sampling grasp poses via a predefined mapping matrix. Raw Image and Raw Image Grid represent two additional end-to-end approaches, both directly processing raw images with CLIP. The latter uses grid-cropped images as input and generates actions through a grasp network. Although LERF-TOGO [23] and GraspSplats [36] achieve fine-grained scene representations through time-consuming training, their grounding accuracy suffers in cluttered settings, leading to cascading errors in action planning. But their performance remains unsatisfactory. Other methods support real-time inference. VLG [37] achieves object awareness by integrating object-centric representations but suffers from detection noise, resulting in low task success rates. ThinkGrasp [38] employs GPT-4o as an object-based crop planner, inheriting the reasoning capabilities of LLMs. However, constrained by segmentation and LLM planning accuracy, it operates in stages, requiring additional planning steps for ambiguous concepts. The A^2 [35] algorithm relies on similarity clouds for grounding, neglecting the probability of moving other obstacles. In contrast, action-priority alignment enables the policy to directly score actions based on task-relevant vision-language features. This approach allows the policy to avoid over-reliance on precise visual representations and overcome obstacles to target grasping. Results in Table 1 demonstrate that the policy outperforms all baselines.

Place the task. Performance in Table 1 demonstrates Place's capabilities with both visible and invisible objects, further highlighting the architecture's advantages. VLP [39] relies heavily on CLIP's capabilities, yet CLIP often fails when confronted with similar visual information or textual words. A^2 strives to distinguish between "in" and "around" relationships by directly establishing the highest point that satisfies both referential and relational requirements, but it ignores 3D layout, leading to unstable placement in cluttered scenes. Decoupling semantic filtering and geometric alignment, it enables robust placement even in complex geometries.

Table 1: Simulation results for all categories and arrangements.

Category	Method	Seen (Success Rate/Planning Steps)	Unseen (Success Rate/Planning Steps)
Grasp	CLIPort-Grounding	59.3/4.8	64.0/4.15
	Raw Image	48.1/4.35	45.3/5.37
	Raw Image Grids	60.7/4.35	70.7/3.49
	LERF-TOGO	83.3/3.37	76.2/2.01
	GraspSplats	58.0/2.05	37.3/1.67
	VLG	74.3/4.11	78.7/3.98
	ThinkGrasp	84.7/2.55	57.3/4.11
	A^2	90.5/2.68	84.2/3.82
	SR(OUR)	92.0/2.34	92.4/2.50
Place	VLP	40.0	20.0
	A^2	66.3	44.4
	SR(OUR)	70.5	48.8
Pick-Place	Act3D	0.0/-	0.0/-
	RVT-2	0.83/4.00	0.0/-
	3D Diffuser Actor	1.67/6.13	0.0/-
	A^2	56.25/3.12	42.5/2.32
	SR(OUR)	60.0/3.45	50.0/3.25

Pick-Place task. As shown in Table 1, RVT-2 [29], Act3D [40], and 3D Diffuser Actor [41] fail in most scenarios when using their pre-trained models for pick-and-place collaborative tasks to test zero-shot generalization, indicating poor generalization capabilities for novel objects, backgrounds, and camera viewpoints. Even when trained on the dataset, RVT-2 [29], Act3D [40], and 3D Diffuser Actor [41] struggle to acquire sufficient information for task completion, which may be due to mismatched data formats or the high degree of clutter. By integrating off-the-shelf foundation priors into a staged semantic–geometric alignment framework guided by zero-shot vision–language representations, our policy improves generalization by leveraging staged alignment, without requiring extensive end-to-end task-specific retraining of the model. All results are summarized in Table 1.

Table 2 lists the statistics for grasp task, grouped in batches of 10 tasks.

Table 2: Single test results.

Lang_goal	Avg_success	Avg_step	Avg_success_step	Avg_reward
I need a fruit	0.9	3.3	2.77778	0.241599
Get something to hold other things	1.0	1.2	1.20000	0.968692
Give me the pear	1.0	3.1	3.10000	0.369404
Give me the thera_med	1.0	5.3	5.30000	0.161902
I need a cup	1.0	1.1	1.10000	0.981099

(Continued)

Table 2 (continued)

Lang_goal	Avg_success	Avg_step	Avg_success_step	Avg_reward
Grasp a round object	1.0	1.1	1.10000	0.835258
Get something to drink	1.0	1.4	1.40000	0.945054
I want a round object	0.9	2.6	2.00000	0.654881
Get something to eat	1.0	1.3	1.30000	0.958561
Give me the cup	0.9	3.7	3.22222	0.467884

4.3 Ablation Experiment

To validate the effectiveness of the proposed method, a comprehensive ablation study was designed and conducted. This experiment aimed to compare the performance of the three variants: a single-stage baseline (BASE, which employs a 2-Align only paradigm directly matching action priors with the global point cloud), a two-stage alignment model without DSSA (SR-A), and the full model (SR) across different task scenarios, particularly their generalization capabilities within the training set (Seen) and test set (Unseen) environments. All learning-based methods were trained under identical experimental settings to ensure fair comparison.

The experimental results are shown in Table 3, which details the average success rates (%) and planning steps for both models across grasping, placing, and combined pick-and-place tasks.

Table 3: Ablation experiment.

	Two-Stage	DSSA	Grasp		Place		Grasp Place	
			Seen	Unseen	Seen	Unseen	Seen	Unseen
BASE			90.7/2.43	80.3/3.21	59.0	38.0	55.20/3.23	40.5/3.66
SR-A	✓		94.1/2.49	76.2/4.02	66.9	32.2	66.25/2.32	37.5/4.37
SR	✓	✓	92.0/2.34	92.4/2.50	70.5	48.8	60.0/3.45	50.0/3.25

Table 3 presents an ablation comparison among BASE, SR-A, SR across grasping, placement, and combined pick-and-place tasks under both Seen and Unseen settings. In Seen environments, BASE and SR-A often achieve comparable or slightly higher success rates, indicating stronger fitting capacity when the test distribution closely matches the training data, with SR-A achieving the highest Seen grasping performance (94.1%). However, this advantage does not generalize to Unseen environments. Under distribution shifts, SR consistently outperforms both BASE and SR-A across all task categories. Importantly, SR maintains success rate and lower planning steps in Unseen grasping and pick-and-place tasks, indicating that the staged semantic–geometric alignment combined with DSSA not only improves robustness but also enables more efficient action planning. Overall, SR trades slight fitting capacity in Seen scenarios for significantly improved generalization and stability in previously unseen and cluttered environments.

We evaluate the robustness of each sparsification function to point cloud noise. We add Gaussian noise to the point cloud with varying intensities (σ = 0.1 for Light, σ = 0.2 for Medium, σ = 0.3 for High). Table 4

reports the average success rates in the unseen grasping task under different noise levels. The results show that performance degrades as noise intensity increases, while $ReLU^2$ consistently outperforms ReLU and softmax across all noise levels, confirming its strong ability to suppress noise interference and retain valid geometric-semantic features, which is critical for point cloud-based alignment tasks.

Table 4: Sensitivity of sparsification functions to point cloud noise.

Sparsification Function	Light Noise	Middle Noise	Heavy Noise
ReLU	84.3	77.1	66.4
Softmax	83.5	74.8	70.2
$ReLU^2$	86.7	78.1	71.3

To investigate the training convergence characteristics of the models, the loss trends of SR-A and SR models were compared across training steps. Experimental results shown in Fig. 8 indicate that both models exhibit similar initial loss values, with both decreasing as training steps increase. Among them, the SR-A model demonstrates a more rapid loss reduction, followed by stable and sustained convergence; the SR model initially decreases at a slightly slower pace, with its subsequent decline gradually moderating before ultimately stabilizing and converging. Notably, the final loss of SR is lower than that of SR-A at the end of training. The lower final loss reflects a higher degree of cross-modal alignment quality and feature selectivity, which directly translates to the substantial performance gains observed in cluttered, unseen environments.

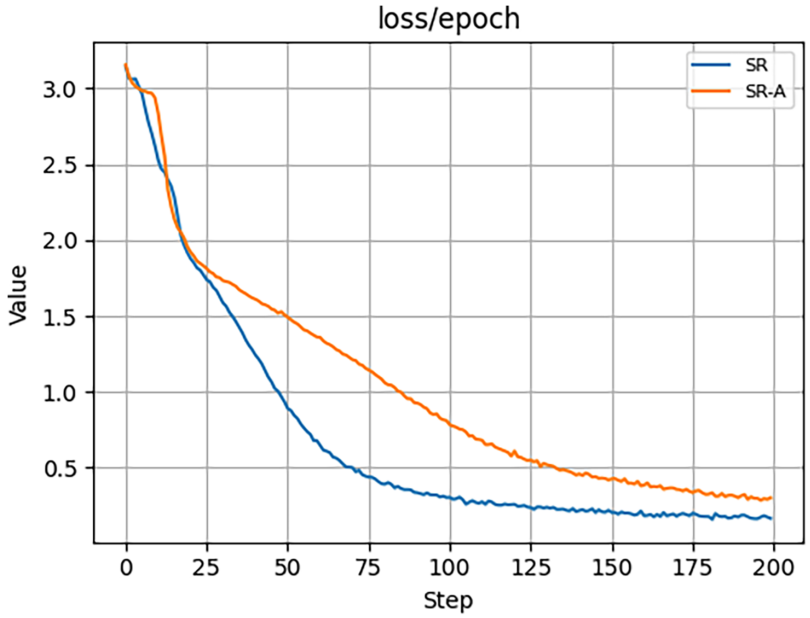

Figure 8: Training loss.

4.4 Case Studies

To visually validate the effectiveness of the proposed method, three typical task scenarios were selected for case studies. The performance of the BASE, SR-A, and SR models was compared, with results shown in the figure.

In the “grasp a round object” task shown in Fig. 9, the BASE model without the 1-Align exhibits significant deviations in target recognition and grasp pose planning, resulting in misidentifications of irrelevant objects. While the SR-A model can localize circular targets, its grasp pose accuracy remains insufficient. In contrast, the SR model, leveraging the DSSA module, precisely focuses on circular targets and plans stable grasp poses, effectively mitigating interference from irrelevant information.

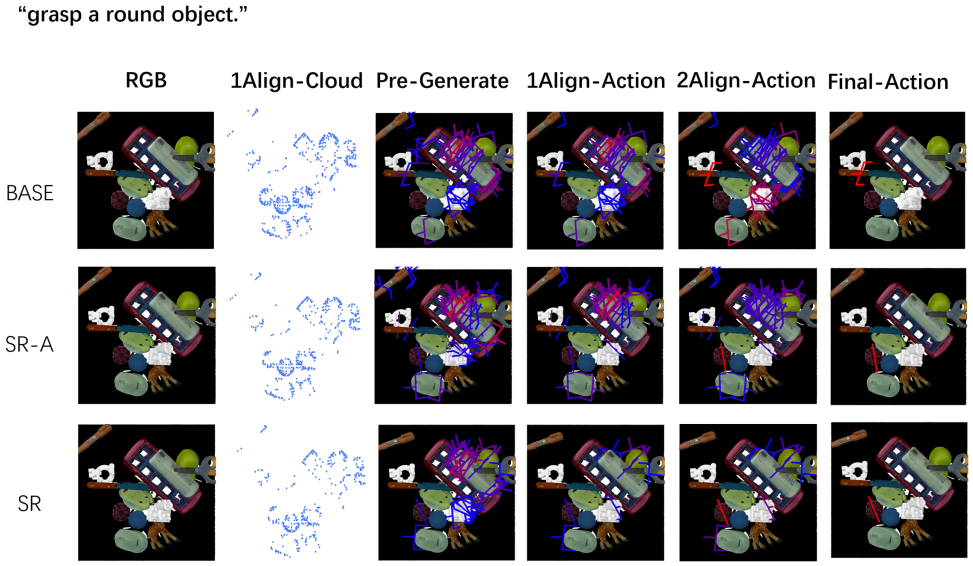

Figure 9: Grasp results comparison.

In summary, the case studies demonstrate that the SR model outperforms both the BASE and SR-A models across multiple embodied intelligence tasks. The introduction of the DSSA module effectively enhances the model’s semantic understanding accuracy and action execution robustness, providing more reliable technical support for robotic operations in complex environments.

5 Conclusion

To address the limitations of previous methods in robotic tasks within complex cluttered environments, particularly insufficient visual grounding accuracy, cumulative errors in module cascading, and weak end-to-end policy generalization, we innovatively propose a Secondary Realignment (SR) framework that integrates visual language models with action foundations. The SR framework achieves semantic alignment through CLIP-driven multimodal 3D scene representations, suppresses irrelevant visual noise, and enhances perception stability in cluttered 3D scenes via DSSA. Extensive PyBullet simulation experiments demonstrate that SR significantly improves success rates over baselines in unseen object scenarios while simultaneously optimizing operational efficiency. Ablation studies rigorously validate the core contributions of the secondary realignment mechanism and DSSA module: the former reduces semantic–geometric ambiguities by decoupling vision–language and vision–action alignment, while the latter enhances perception stability in complex environments. Acknowledging that current validations are restricted to simulation, future work will focus on extending SR to real-robot setups for quantitative physical validation. Additionally, we plan to deepen multimodal pre-training to improve semantic understanding and adapt the framework to dynamic stacked scenes and human-robot collaborative interference scenarios, aiming to achieve practical deployment in industrial and service settings.

Acknowledgement: Not applicable.

Funding Statement: Fujian Provincial Science and Technology Department: Collaborative Innovation Platform Project for Key Technologies of Smart Warehousing and Logistics System in Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone (2025E3024). Key Technical Innovation and Industrialization Project of Fujian Province's Manufacturing Industry in 2025: Research and Development and Industrialization of Key Technologies for Ultra-Low Power Consumption Multi-Modal Intelligent Sensing Terminals and Anti-Interference AI Algorithms (2025G006).

Author Contributions: Study conception and design: Jinjiang Lin; data collection: Yuan Lu and Han Li; analysis and interpretation of results: Xiaolong Cai and Enyi Chen; draft manuscript preparation: Jiansheng Guan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu Z, Zhou Y, Xu X, Wang Z, Yan H. MoManipVLA: transferring vision-language-action models for general mobile manipulation. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 1714–23. doi:10.1109/CVPR52734.2025.00167.
2. Zhu J, Sun X, Zhang Q, Liu M. VLA-Grasp: a vision-language-action modeling with cross-modality fusion for task-oriented grasping. Complex Intell Syst. 2025;11(6):272. doi:10.1007/s40747-025-01893-x.
3. Li S, Wang J, Dai R, Ma W, Ng WY, Hu Y, et al. RoboNurse-VLA: robotic scrub nurse system based on vision-language-action model. In: Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2025 Oct 19–25; Hangzhou, China. p. 3986–93. doi:10.1109/IROS60139.2025.11246030.
4. Xu J, Jin S, Lei Y, Zhang Y, Zhang L. RT-grasp: reasoning tuning robotic grasping via multi-modal large language model. In: Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2024 Oct 14–18; Abu Dhabi, United Arab Emirates. p. 7323–30. doi:10.1109/IROS58592.2024.10801718.
5. Mei A, Zhu GN, Zhang H, Gan Z. RePlanVLM: replanning robotic tasks with visual language models. IEEE Robot Autom Lett. 2024;9(11):10201–8. doi:10.1109/LRA.2024.3471457.
6. Jeong H, Lee H, Kim C, Shin S. A survey of robot intelligence with large language models. Appl Sci. 2024;14(19):8868. doi:10.3390/app14198868.
7. Huang J, Limberg C, Arshad SMN, Zhang Q, Li Q. Combining VLM and LLM for enhanced semantic object perception in robotic handover tasks. In: Proceedings of the 2024 WRC Symposium on Advanced Robotics and Automation (WRC SARA); 2024 Aug 23; Beijing, China. p. 135–40. doi:10.1109/WRCSARA64167.2024.10685688.
8. Zhou Q, Gu Y, Li J, Feng B, Li B, Bi Y. Towards zero-shot robot tool manipulation in industrial context: a modular VLM framework enhanced by multimodal affordance representation. Robot Comput Integr Manuf. 2026;98(2):103161. doi:10.1016/j.rcim.2025.103161.
9. Xu K, Yu H, Lai Q, Wang Y, Xiong R. Efficient learning of goal-oriented push-grasping synergy in clutter. IEEE Robot Autom Lett. 2021;6(4):6337–44. doi:10.1109/LRA.2021.3092640.
10. Myers A, Teo CL, Fermüller C, Aloimonos Y. Affordance detection of tool parts from geometric features. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA); 2015 May 26–30; Seattle, WA, USA. p. 1374–81. doi:10.1109/ICRA.2015.7139369.
11. Song D, Ek CH, Huebner K, Kragic D. Task-based robot grasp planning using probabilistic inference. IEEE Trans Robot. 2015;31(3):546–61. doi:10.1109/TRO.2015.2409912.

12. Chu FJ, Xu R, Vela PA. Learning affordance segmentation for real-world robotic manipulation via synthetic images. IEEE Robot Autom Lett. 2019;4(2):1140–7. doi:10.1109/LRA.2019.2894439.
13. Li S, Bhagat S, Campbell J, Xie Y, Kim W, Sycara K, et al. ShapeGrasp: zero-shot task-oriented grasping with large language models through geometric decomposition. In: Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2024 Oct 14–18; Abu Dhabi, United Arab Emirates. p. 10527–34. doi:10.1109/IROS58592.2024.10801661.
14. Bohg J, Morales A, Asfour T, Kragic D. Data-driven grasp synthesis—a survey. IEEE Trans Robot. 2014;30 (2):289–309. doi:10.1109/TRO.2013.2289018.
15. Zhang H, Peeters J, Demeester E, Kellens K. A CNN-based grasp planning method for random picking of unknown objects with a vacuum gripper. J Intell Rob Syst. 2021;103(4):64. doi:10.1007/s10846-021-01518-8.
16. Kumra S, Joshi S, Sahin F. GR-ConvNet v2: a real-time multi-grasp detection network for robotic grasping. Sensors. 2022;22(16):6208. doi:10.3390/s22166208.
17. Ma H, Qin R, Shi M, Gao B, Huang D. Sim-to-real grasp detection with global-to-local RGB-D adaptation. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 13910–7. doi:10.1109/ICRA57147.2024.10611165.
18. Mahler J, Goldberg K. Learning deep policies for robot Bin picking by simulating robust grasping sequences. In: Proceedings of the Conference on Robot Learning; 2017 Nov 13–15; Mountain View, CA, USA.
19. Satish V, Mahler J, Goldberg K. On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks. IEEE Robot Autom Lett. 2019;4(2):1357–64. doi:10.1109/LRA.2019.2895878.
20. Li J, Cappelleri DJ. Sim-grasp: learning 6-DOF grasp policies for cluttered environments using a synthetic benchmark. IEEE Robot Autom Lett. 2024;9(9):7645–52. doi:10.1109/LRA.2024.3430712.
21. Song Y, Sun P, Jin P, Ren Y, Zheng Y, Li Z, et al. Learning 6-DoF fine-grained grasp detection based on part affordance grounding. IEEE Trans Automat Sci Eng. 2025;22(2):15200–14. doi:10.1109/tase.2025.3566461.
22. Xiang J, Tao T, Gu Y, Shu T, Wang Z, Yang Z, et al. Language models meet world models: embodied experiences enhance language models. Adv Neural Inf Process Syst. 2023;36:75392–412. doi:10.52202/075280-3295.
23. Rashid A, Sharma S, Kim CM, Kerr J, Chen LY, Kanazawa A, et al. Language embedded radiance fields for zero-shot task-oriented grasping. In: Proceedings of the 7th Annual Conference on Robot Learning; 2023 Nov 6–9; Atlanta, GA, USA. p. 178–200.
24. Chi C, Xu Z, Feng S, Cousineau E, Du Y, Burchfiel B, et al. Diffusion policy: visuomotor policy learning via action diffusion. Int J Robot Res. 2025;44(10–11):1684–704. doi:10.1177/02783649241273668.
25. Lin YC, Zeng A, Song S, Isola P, Lin TY. Learning to see before learning to act: visual pre-training for manipulation. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France. p. 7286–93. doi:10.1109/ICRA40945.2020.9197331.
26. Hoang DC, Nguyen AN, Vu VD, Vu DQ, Nguyen VT, Nguyen TU, et al. Grasp configuration synthesis from 3D point clouds with attention mechanism. J Intell Robot Syst. 2023;109(3):71. doi:10.1007/s10846-023-02007-w.
27. Fang HS, Wang C, Fang H, Gou M, Liu J, Yan H, et al. AnyGrasp: robust and efficient grasp perception in spatial and temporal domains. IEEE Trans Robot. 2023;39(5):3929–45. doi:10.1109/TRO.2023.3281153.
28. James S, Wada K, Laidlow T, Davison AJ. Coarse-to-fine Q-attention: efficient learning for visual robotic manipulation via discretisation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 13729–38. doi:10.1109/CVPR52688.2022.01337.
29. Goyal A, Blukis V, Xu J, Guo Y, Chao YW, Fox D. RVT-2: learning precise manipulation from few demonstrations. In: Proceedings of the Robotics: Science and Systems 2024; 2024 Jul 15–19; Delft, The Netherlands.
30. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
31. Fang HS, Wang C, Gou M, Lu C. GraspNet-1Billion: a large-scale benchmark for general object grasping. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11441–50. doi:10.1109/cvpr42600.2020.01146.

32. Zhang B, Titov I, Sennrich R. Sparse attention with linear units. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Punta Cana, Dominican Republic. p. 6507–20.
33. Coumans E, Bai Y. PyBullet, a Python module for physics simulation for games, robotics and machine learning 2016 [Internet]. [cited 2026 Mar 9]. Available from: http://pybullet.org.
34. Shridhar M, Manuelli L, Fox D. CLIPort: what and where pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning; 2021 Nov 8–11; London, UK. p. 894–906.
35. Xu K, Xia X, Wang K, Yang Y, Mao Y, Deng B, et al. Efficient alignment of unconditioned action prior for language-conditioned pick and place in clutter. IEEE Trans Automat Sci Eng. 2025;22:21256–68. doi:10.1109/tase.2025.3606549.
36. Ji M, Qiu R, Zou X, Wang X. GraspSplats: efficient manipulation with 3D feature splatting. In: Proceedings of the 8th Conference on Robot Learning; 2024 Nov 6–9; Munich, Germany. p. 1443–60.
37. Xu K, Zhao S, Zhou Z, Li Z, Pi H, Zhu Y, et al. A joint modeling of vision-language-action for target-oriented grasping in clutter. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK. p. 11597–604. doi:10.1109/ICRA48891.2023.10161041.
38. Qian Y, Zhu X, Biza O, Jiang S, Zhao L, Huang H, et al. ThinkGrasp: a vision-language system for strategic part grasping in clutter. In: Proceedings of the 8th Conference on Robot Learning; 2024 Nov 6–9; Munich, Germany. p. 3568–86.
39. Xu Z, Xu K, Xiong R, Wang Y. Object-centric inference for language conditioned placement: a foundation model based approach. In: Proceedings of the 2023 International Conference on Advanced Robotics and Mechatronics (ICARM); 2023 Jul 8–10; Sanya, China. p. 203–8. doi:10.1109/ICARM58088.2023.10218865.
40. Gervet T, Xian Z, Gkanatsios N, Fragkiadaki K. Act3D: 3D feature field transformers for multi-task robotic manipulation. In: Proceedings of the 7th Conference on Robot Learning; 2023 Nov 6–9; Atlanta, GA, USA. p. 3949–65.
41. Ke TW, Gkanatsios N, Fragkiadaki K. 3D diffuser actor: policy diffusion with 3D scene representations. In: Proceedings of the 8th Conference on Robot Learning; 2024 Nov 6–9; Munich, Germany. p. 1949–74.