# GraspMamba: A Mamba-based Language-driven Grasp Detection Framework with Hierarchical Feature Learning

Huy Hoang Nguyen[2], An Vuong[3], Anh Nguyen[4], Ian Reid[3], Minh Nhat Vu[1,2]

***Abstract*— Grasp detection is a fundamental robotic task critical to the success of many industrial applications. However, current language-driven models for this task often struggle with cluttered images, lengthy textual descriptions, or slow inference speed. We introduce GraspMamba, a new language-driven grasp detection method that employs hierarchical feature fusion with Mamba vision to tackle these challenges. By leveraging rich visual features of the Mamba-based backbone alongside textual information, our approach effectively enhances the fusion of multimodal features. GraspMamba represents the first Mamba-based grasp detection model to extract vision and language features at multiple scales, delivering robust performance and rapid inference time. Intensive experiments show that GraspMamba outperforms recent methods by a clear margin. We validate our approach through real-world robotic experiments, highlighting its fast inference speed.**

## I. INTRODUCTION

Robotic grasping is an important task with several applications and has received growing attention from researchers in the past decades [1]–[3]. Traditional grasp detection methods [4] often overlook the use of language prompts [5], limiting the agent's ability to interpret language instructions beyond their literal meaning and resulting in unpredictable behavior [6]. Emerging language-driven grasp methodologies [7], [8] have enabled robots to grasp specific objects based on language prompts [9]. These robotic systems, trained on large-scale datasets, provide robust visual-language understanding, like determining which object part to grasp based on human instructions [10], showing promise in developing general-purpose robots [11].

Recently, language-driven grasping has gained attention as a promising research area in robotic manipulation. The use of language can be viewed as a representation of the scenario surrounding objects, conveying information to humans or machines for conceptual understanding. For example, by giving a command to "grasp a cup on the table," the robot knows where a cup is and can determine the specific grasp actions for objects. Therefore, by leveraging language, many studies attempt to bridge the gap between vision and language for robotic applications [9], [12]–[14]. For example, Roco [15] and Manipllm [16] are robotic language models designed to provide instructions for robots operating in real-world environments by leveraging large language models such as [17] or large vision language models [18]–[20]. Previous studies have examined multiple approaches for integrating language understanding into robotic grasp detection. One approach treats this as a grasp-pose generation problem conditioned on language prompts [21] or uses diffusion models [9] to achieve promising results. However, diffusion models face challenges in real-time robotics due to their long inference times [13]. Alternatively, methods that leverage transformers [12], [18], [22] successfully combine textual and visual features but often struggle with object complexities [23]. This limitation is particularly evident in scenarios requiring long-range visual-language dependencies and handling images with extensive textual descriptions, which restricts their application to real-world robots.

Recently, Mamba [24], with its global receptive field coverage and dynamic weights offering linear complexity [25], presents an ideal solution for addressing long-range visual-language dependencies while maintaining fast inference speeds, making it well-suited for robotic applications [26]. Mamba has demonstrated exceptional effectiveness in tasks involving long sequence modeling, especially in natural language processing [27]. Researchers have begun exploring its potential for vision-related applications, including image classification [28]–[30], image segmentation [31]–[33], and point cloud analysis [25], [34], [35]. Roboticists are also investigating how Mamba's context-aware reasoning and linear complexity can be applied to solve robotic tasks [26], [36]. In line with this direction, this paper aims to leverage Mamba to integrate image and text modalities to generate semantically plausible grasping poses for robotic systems.

This paper introduces **GraspMamba**, the first language-driven grasp detection framework built on the State Space Model. Specifically, we propose a novel hierarchical feature fusion technique that integrates textual features with visual features at each stage of the hierarchical vision backbone. We argue that our Mamba-based fusion technique addresses the computational inefficiencies of transformers caused by the doubled sequence length when combining text and vision features [37]. Our method maintains strong performance in grasp detection by efficiently learning spatial information and hierarchically incorporating textual features to enhance context and semantics. By aligning textual features within a shared space, the model effectively merges multimodal representations across multiple scales, which is known for enabling grounding objects [38]. We validate our approach on a recent large-scale language-driven grasping dataset [9], demonstrating better accuracy and faster inference than current state-of-the-art methods. Additionally, our approach supports zero-shot learning and is generalized to real-world robotic grasping applications.

[1] Automation & Control Institute, TU Wien, Austria
[2] AIT Austrian Institute of Technology GmbH, Austria
[3] Department of Computer Vision, MBZUAI, UAE
[4] Department of Computer Science, University of Liverpool, UK

Our main contributions are summarized as follows:

- We propose a novel vision-language model built upon the Mamba technique, designed to fuse vision and language features within a hidden state space. This approach establishes a novel framework for integrating multimodal information, improving efficiency and accuracy in language-driven grasp detection.
- We validate **GraspMamba** 's applicability by showing an in-depth analysis of our proposed method and presenting experimental results on benchmark datasets. Our findings demonstrate that it surpasses other approaches in accuracy and execution speed. Our code and models will be released.

## II. Related Work

**Grasp Detection.** Traditional robot grasp detection methods include analytic approaches [39], [40] that rely on kinematic and dynamic models to identify stable grasp points, ensuring they meet flexibility, balance, and stability criteria. In contrast, several data-driven approaches based on machine learning [41]–[44] have been developed to enable robots to learn and mimic human grasping strategies through deep learning techniques. These approaches have been further enhanced by the use of RGB-D images [45], [46] and 3D point clouds [47], [48], allowing for grasp detection in 3D space. However, one major limitation of both analytic and CNN-based methods is their restricted scene understanding and inability to process language instructions, which reduces their effectiveness in dynamic, human-centered environments.

**Language-driven Grasping**. Language-driven grasping represents the use of natural language to localize the object region for grasping [9], [13], [49]–[55]. Recent research focuses on establishing correlations between textual embeddings and vision embeddings within a shared embedding space. This approach aims to identify the target object and subsequently generate the grasping pose based on the foundation models [14]. Bhat *et al.* [56] introduce a method that fuses image and text embeddings by leveraging the lightweight segmentation decoder. However, these methods are hindered by high computational and memory demands during training and inference.

**Cross-Modal Feature Fusion.** Multimodal feature fusion is a critical technique in various applications to optimize the alignment between linguistic and visual domains. Conventionally, many approaches have focused on independently mapping global features of images and sentences into a shared embedding space to compute image-sentence similarity [57]–[60]. Recent advancements, such as the work by Xu *et al.* [61], capture higher-order interactions between visual regions and textual elements by incorporating inter- and intra-modality relations in the feature fusion process. Furthermore, the attention mechanism in Transformers [62] and cosine similarity metrics have been widely used to enhance the alignment of textual and visual embeddings, improving multimodal representation learning. However, these multimodal feature fusion methods face limitations when focusing on fine-grained image regions or when processing queries with limited textual information or complex sentences.

**State Space Models for Vision-and-Language.** State Space Models (SSMs) with selection mechanisms and hardware-aware architectures have recently demonstrated substantial promise in long-sequence modeling. The original SSM block is designed for processing one-dimensional sequences, while vision-related tasks necessitate handling multi-dimensional inputs like images, videos, and 3D representations. Several approaches have been proposed to adapt SSMs for complex vision-related applications. For instance, ViM [63], also called the Bidirectional Mamba block, annotates image sequences with position embeddings and condenses visual representations using bidirectional state space models. Additionally, PlainMamba [64] and EfficientVMamba [65] improve the capabilities of visual state space [66] blocks by stacking multiple blocks on the feature map and employing different scanning approaches. While these methods effectively address the need for global context and spatial understanding, their increased complexity can lead to challenges in training and a higher risk of overfitting. To mitigate these issues, Hatamizadeh *et al.* [67] introduced a hybrid Mamba-Transformer backbone that enhances global context representation learning.

Despite Mamba's increasing popularity in vision tasks [34], [68], there remains a significant gap in integrating text and image modalities [69], [70]. A key challenge in vision-and-language Mamba models is using a single projection from the image to the language domain [71], which fails to capture image features at multiple resolutions [38]. To mitigate this issue, we propose a hierarchical feature fusion method that integrates vision and text features at various scales, leveraging Mamba's efficient computation [67]. Specifically, we integrate rich textual information from a text encoder at each stage of the vision backbone to enhance global multimodal information, retaining all crucial features through an element-wise technique. Consequently, the output for grasp detection preserves the essential information from the input data at multiple scales, which can serve as robust guidance to solve such fine-grained generative problems [72] like the language-driven grasp generation. Experimental results confirm that our Mamba-based fusion method delivers competitive performance with faster, linearly scalable inference and constant memory usage in both vision and robotic applications.

## III. GraspMamba

### A. Overview

We propose a method for detecting the grasping pose of an object by integrating textual features with rich visual features derived from a Mamba-based architecture. Given an input RGB image and a corresponding text prompt describing the object of interest, our approach aims to accurately identify the object's grasping pose. Following the established convention of *rectangle grasp*, as outlined in [73], we define each grasping pose using five parameters: the center coordinates

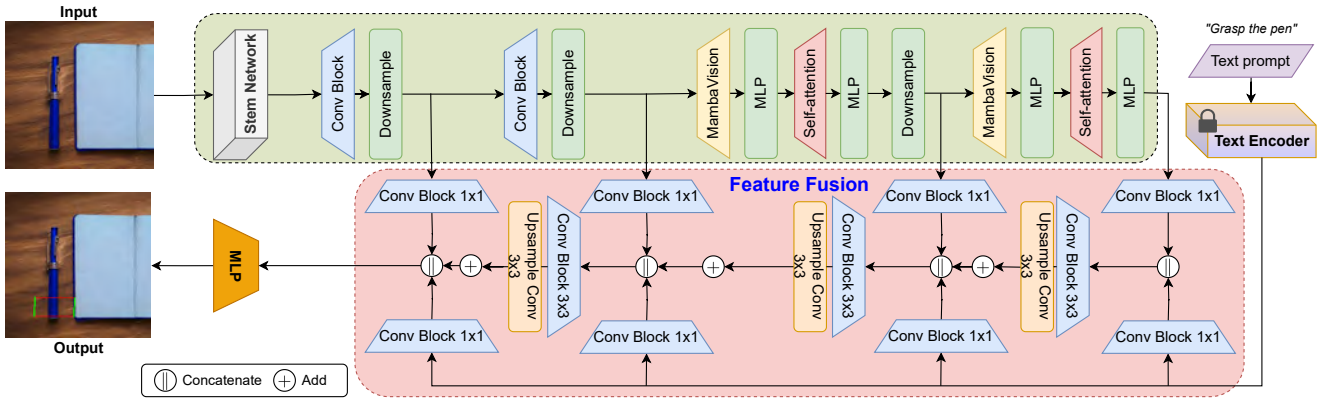



Fig. 1. The overview of our **GraspMamba** framework for the language-driven grasp detection task.

$(x, y)$, the rectangle's width and height $(w, h)$, and the rotational angle that indicates the rectangle's orientation relative to the image's horizontal axis. The overall framework of our method is depicted in Fig. 1.

### B. Visual and Language Feature Extraction

**Mamba Visual Feature Extraction.** Inspired by the powerful Swin Transformer [38] hierarchical design, we adopt a four-stage structure to balance speed and accuracy in vision tasks. Therefore, we leverage weights of pre-trained MambaVision [67] to produce multi-level representations at each stage. Given an image of size $H \times W \times 3$, the initial two stages consist of CNN-based layers for fast feature extraction at higher input resolutions with size of $\frac{H}{4} \times \frac{W}{4} \times C$ and $\frac{H}{8} \times \frac{W}{8} \times 2C$, respectively. Subsequently, the MambaVision and multihead self-attention [62] blocks are applied afterward as referred to as feature transformation stages, with output resolution of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively. Specifically, the MambaVision block modifies the original Mamba by creating the symmetric path without SSM as a token mixer to enhance the modeling of the global context. SSMs map one-dimensional sequence $\mathbf{x}(t) \in \mathbb{R}^L$ to $\mathbf{y}(t) \in \mathbb{R}^L$ through a hidden state $\mathbf{h}(t) \in \mathbb{R}^N$. With the evolution parameter $\mathbf{A} \in \mathbb{R}^{N \times N}$ and the projection parameters $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$, such a model is formulated as linear ordinary differential equations:

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{1a}$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t). \tag{1b}$$

As continuous-time models, state space models are adapted for deep learning applications with discrete data space through a discretization step using Zero-Order Hold assumption [24]. In this transformation, the continuous-time parameters $\mathbf{A}$ and $\mathbf{B}$ are converted into their discrete-time equivalents, denoted as $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, respectively, with a timescale parameter $\Delta$ according to:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{2a}$$
$$\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \tag{2b}$$

Thus, Equation (1) can be rewritten as:

$$\mathbf{h}_t = \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}\mathbf{x}_t, \tag{3a}$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t. \tag{3b}$$

To improve computational performance and allow for better scaling, the iterative process in Equation (3) can be synthesized through a global convolution

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \cdots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \tag{4a}$$
$$\mathbf{y} = \mathbf{x} * \overline{\mathbf{K}}, \tag{4b}$$

where $L$ is the length of the input sequence $\mathbf{x}$, $\overline{\mathbf{K}} \in \mathbb{R}^L$ serves as the kernel of the SSMs and $*$ represents the convolution operation. Using hierarchical representations at different stages of the Mamba-based backbone, our method effectively captures global structures and fine-grained details, enhancing performance across vision-related tasks.

**Text Embedding.** Following the standard practice [9], we encode the input query (e.g., "grasp a pencil") using a pre-trained CLIP [74], and SigLIP [75] model, producing text embedding features $\mathbf{T} \in \mathbb{R}^{B \times C_\mathbf{T}}$.

### C. Hierarchical Feature Fusion

While Mamba is effective in modeling long sequences, many Mamba-based multimodal approaches treat multimodal data as a single domain sequence rather than focusing on how to integrate features effectively [68], [76], [77]. To address this, inspired by the hierarchical nature of Swin Transformer [38], we aim to develop a new approach to fuse visual and textual features in a multiscale manner. Additionally, our hierarchical feature fusion is a simple, learnable module that aligns the vision and text features by transforming the dimensionality of the textual representation to match the token dimensions used in the Mamba-based model to create rich, multimodal representation for tasks such as visual grounding or language-driven grasp detection. Our hierarchical feature fusion block is shown in Fig. 1.

To combine visual and textual information effectively, we begin by aligning their dimensions and preparing them for fusion. This process involves applying $1 \times 1$ convolutions to both the image and the text features, reducing their channel

dimensions to a common space while allowing the network to learn combined feature representations for fusion. We then expand the text features to match the spatial dimensions of the image features, ensuring that each spatial location in the image can attend to the entire text representation. These processed features are then concatenated along the channel dimension, creating a unified representation that preserves information from both modalities. Let $\mathbf{X}_l \in \mathbb{R}^{B\times C_I\times H_l\times W_l}$ represent the image features at level $l \in \{1,\dots,L\}$, $\mathbf{T} \in \mathbb{R}^{B\times C_\mathbf{T}}$ represent the text features, and $\mathbf{T}_{\text{exp}}$ represent the expanded text features to match the spatial dimensions of $\mathbf{X}_l$. The visual-language fusion at level $l$, denoted as $\Phi_l$, is defined as:

$$Z_l = \text{Concat}\left(\text{Conv}^{\mathbf{X}}_{1\times1}(\mathbf{X}_l), \text{Conv}^{\mathbf{T}}_{1\times1}(\mathbf{T}_{\text{exp}})\right) \tag{5}$$

To further integrate the concatenated features and capture spatial relationships, we apply a 3x3 convolution. This crucial step allows for local feature interactions between the image and text modalities, helps to learn spatially-aware multimodal representations, and increases the receptive field to capture more context. This integration is achieved through:

$$\Phi_l(\mathbf{X}_l, \mathbf{T}) = \text{Conv}_{3\times3}\left(Z_l\right) \tag{6}$$

where $\text{Conv}^{\mathbf{X}}_{1\times1}$ and $\text{Conv}^{\mathbf{T}}_{1\times1}$ are $1\times1$ convolutions applied to the image and text features, respectively. $\text{Conv}_{3\times3}$ is a $3\times3$ convolution.

To preserve global information across different levels of the vision backbone, we define an upscaling operation $U_l$, which allows information to flow from deeper layers to shallower layers, helps to preserve fine-grained details while incorporating global context, and enables the model to make more informed decisions at each level of the hierarchy. The upscaling operation is applied to the hierarchical feature fusion $F_l$ in layer $l$ is defined as:

$$U_l(F_l) = \text{BilinearUpsample}\left(\text{Conv}_{3\times3}(F_l)\right) \tag{7}$$

Here, $F_l$ represents the hierarchical feature fusion in layer $l$, which is introduced recursively to capture and refine multiscale information across the network. By adding $U_{l+1}(F_{l+1})$ at layer $l$, we merge high-level semantic cues from deeper layers with the local, level-specific details at shallower layers. Specifically, the upscaling operation $U_{l+1}$ reshapes the features of the deeper layer $F_{l+1}$ to match the spatial resolution of the layer $l$, allowing them to be added to $\Phi_l(\mathbf{X}_l, \mathbf{T})$. This recursive property is essential as it enables the model to capture multiscale information effectively and facilitates the integration of global and local features at each level. Hierarchical feature fusion is recursively defined as:

$$F_l = \begin{cases} \Phi_L(\mathbf{X}_L, \mathbf{T}), & \text{if } l = L, \\ \Phi_l(\mathbf{X}_l, \mathbf{T}) + U_{l+1}(F_{l+1}), & \text{if } 1 \le l < L. \end{cases} \tag{8}$$

Inspired by GR-ConvNet [78], the final high-dimensional features $F_l$ are subsequently transformed into the grasp detection output through a composition of multiple MLP layers.

## IV. Experimental Results

The experiments initially focus on evaluating the effectiveness of our approach on the Grasp-Anything dataset [7]. We test our proposed method on real robot grasp detection tasks. Additionally, we conduct ablation studies to analyze our process in the context of language-driven grasp detection. Finally, we discuss the challenges and highlight open questions for future research.

### A. Experimental Setup

**Dataset.** To assess the generalization of all methods, we set up our experiments by training on the Grasp-Anything dataset [7]. This dataset is created from large-scale foundation models that offer 1M images with textual descriptions. Following the approaches in [7], [79], we split the data into 'Seen' and 'Unseen' categories, designating 70% of the categories as 'Seen' and the remaining 30% as 'Unseen.' We also employ the harmonic mean ('H') metric to assess overall success rates [79]. In addition to the dataset details, our training is conducted over 50 epochs, with 1000 batches per epoch and a batch size of 8. We employ the Adam optimizer with a base learning rate of 1e-3, using a linear warmup over the first 500 batches and applying gradient clipping (max norm of 1.0) to ensure stable training.

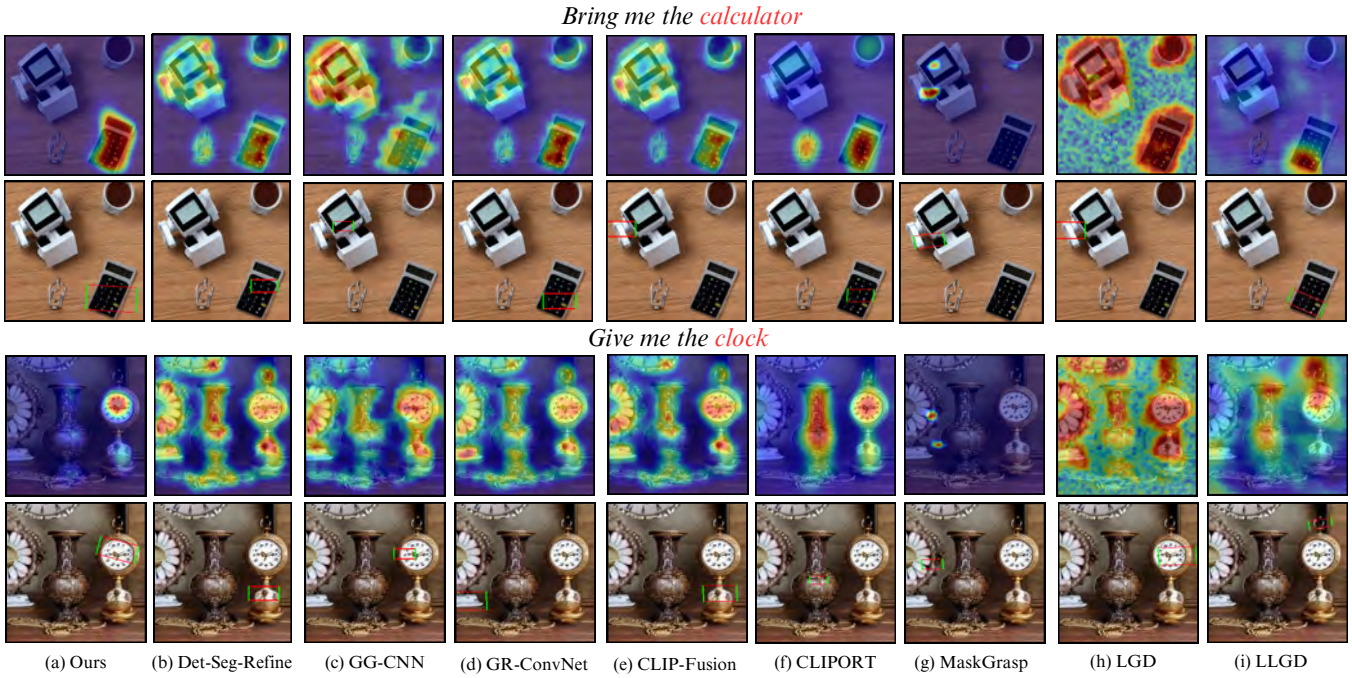
**Evaluation Metrics.** Our primary evaluation metric is the success rate, defined similarly to [78]. A grasp is considered successful if the Intersection over Union (IoU) score between the predicted grasp and the ground truth exceeds 25%, and the offset angle is less than 30 degrees. The text encoder is frozen during training, while the pre-trained vision backbone is fine-tuned using our dataset. We also measure the inference time of all methods using the same NVIDIA RTX 4080 GPU, Intel i7 12700K.

**Baselines.** We compare our method GraspMamba with recent state-of-the-art language-driven grasp detection methods, including: GR-CNN [78], Det-Seg-Refine [80], GG-CNN [81], CLIP-Fusion [53], MaskGrasp [12], LGD [9], LLGD [13], GraspSAM [14] and CLIPORT [18], utilizing pretrained CLIP [74], and SigLIP [75] models for text embedding. **Bold** and underline mean the best result and second best result respectively.

TABLE I

Language-driven grasp detection results.

| Baseline | Seen | UnSeen | H | Inference time |
|---|---|---|---|---|
| Det-Seg-Refine [80] + CLIP [74] | 0.30 | 0.15 | 0.20 | 0.200s |
| GG-CNN [81] + CLIP [74] | 0.12 | 0.08 | 0.10 | 0.040s |
| GR-ConvNet [78] + CLIP [74] | 0.37 | 0.18 | 0.24 | **0.022s** |
| CLIP-Fusion [53] | 0.40 | 0.29 | 0.33 | 0.157s |
| CLIPORT [18] | 0.36 | 0.26 | 0.29 | 0.131s |
| LGD [9] | 0.48 | 0.42 | 0.45 | 22.00s |
| MaskGrasp [12] | 0.50 | 0.46 | 0.45 | 0.116s |
| LLDG [13] | 0.53 | 0.39 | 0.46 | 0.264s |
| GraspSAM [14] | 0.64 | **0.62** | **0.63** | 0.510s |
| GraspMamba + CLIP [74] (ours) | 0.69 | 0.42 | 0.56 | 0.030s |
| GraspMamba + SigLIP [75] (ours) | **0.73** | 0.44 | 0.59 | 0.029s |




(a) Ours (b) Det-Seg-Refine (c) GG-CNN (d) GR-ConvNet (e) CLIP-Fusion (f) CLIPORT (g) MaskGrasp (h) LGD (i) LLGD

Fig. 2. Visual language-driven grasp detection results on examples from the 'Seen' testing set of different methods.

### B. Language-driven Grasp Detection Results

**Quantitative Results.** Table I summarises the results of all methods. This table shows that our GraspMamba significantly outperforms other grasp detection techniques on the Grasp-Anything dataset. Our method consistently achieves better results than other baseline approaches in the 'Seen' scenario. Our method substantially improves, surpassing other methods in the "Seen" scenario. Additionally, our inference time remains competitive, thanks to its simple architecture. Compared to diffusion-based methods [9], [13], our GraspMamba demonstrates a good balance between accuracy and inference speed. We further observe that the variant incorporating SigLIP outperforms the one using CLIP. The key difference is in the loss functions: while CLIP uses a softmax loss that normalizes across the entire batch and pits all negative examples against each other, SigLIP applies a pairwise sigmoid loss that treats each image–text pair individually. This approach reduces the negative impact of semantically similar negatives, a significant advantage for text embeddings that capture subtle nuances. While our method excels on 'Seen' objects, its performance on 'Unseen' objects is lower than the GraspSAM baseline. We note that GraspSAM [14] performs well on unseen cases as it was fine-tuned from a foundation model (Segment Anything [82]), which utilizes extra extensive data for training, while our method is trained from scratch. Furthermore, our inference time is approximately 15 times faster than GraspSAM [14].

**Qualitative Results.** Fig. 2 shows the quantitative evaluation of our method and other baselines. This figure shows that our method produces semantically plausible results, particularly in cluttered scenes with occlusions. The attention maps of our method also show an accurate connection between the text prompt and the visual region.

TABLE II
FEATURE FUSION ANALYSIS.

| Baseline | Seen | UnSeen | H |
|---|---|---|---|
| **GraspMamba** without feature fusion | 0.66 | 0.40 | 0.53 |
| **GraspMamba** with feature fusion | **0.73** | **0.44** | **0.59** |

### C. Ablation Study

**Hierarchical Feature Fusion Analysis.** To evaluate the performance of our hierarchical feature fusion block, we experiment with and without using the feature fusion block in our architecture. Table II summarises the results. We can see that the hierarchical feature fusion block positively impacts the grasp detection performance. Additionally, we visualize the attention maps generated by different methods using the text command. As shown in the heatmap row of Fig. 2, our approach effectively concentrates attention on the target object with minimal distraction from the surrounding area, distinguishing it from other methods. Our feature fusion technique successfully directs the model's focus toward essential regions, enabling the extraction of richer contextual information and enhancing grasp accuracy. While other methods tend to be less precise, with more background interference and a broader focus area. In addition, Fig. 3 indicates our method's effectiveness in aligning visual features with textual inputs. Overall, our method can locate and understand the referenced object based on textual instructions under the variability in different textual instructions while maintaining consistent visual alignment.

### D. Qualitative Analysis

**In the Wild Detection.** Fig. 4 presents in the wild visualization results by our method, which is exclusively

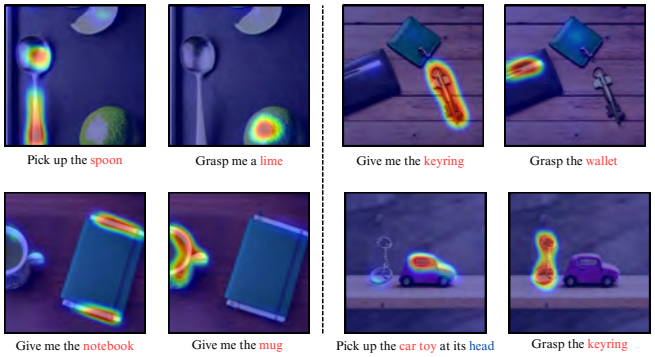



Fig. 3. The visual feature fusion results on examples from the 'Seen' (left) and 'UnSeen' (right) testing set when different text inputs are used.

trained on the Grasp-Anything dataset. These examples, applied to various random internet images and images from other datasets, illustrate our model's strong ability to generalize to real-world scenarios, even though it was trained entirely on synthetic data from the Grasp-Anything dataset. This generalization is critical for deploying grasp detection systems in real-world applications, where the complexity and variability of environments often differ from training datasets.

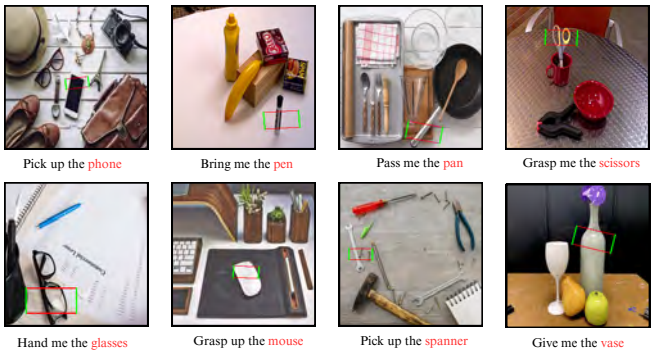



Fig. 4. In the wild detection results. The images are from the YCB-Video [83] dataset and the internet.

**Failure Cases.** Despite successfully improving the alignment of textual and visual features, our method still produces incorrect grasping poses in some cases. The wide variety of objects and grasping prompts presents a significant challenge as the network struggles to capture the full range of real-life scenarios. Fig. 5 illustrates several failure cases where our GraspMamba makes incorrect predictions. Sometimes, the text prompt is aligned with the correct object, but the grasping pose remains inaccurate. This can be attributed to multiple similar objects or shapes that are difficult to differentiate and text prompts lacking sufficient detail for precise predictions.


Fig. 5. Failure cases of our method on examples from the 'Unseen' testing set and the internet.

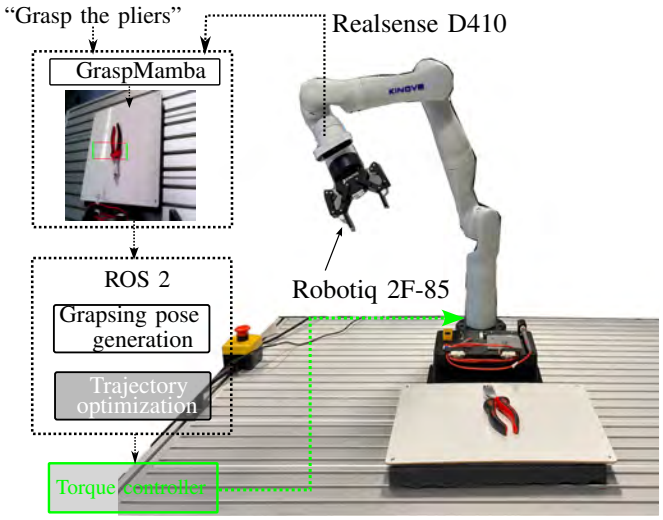



Fig. 6. The robotic experiment setup.

TABLE III

ROBOTIC LANGUAGE-DRIVEN GRASP DETECTION RESULTS

| Baseline | Single | Cluttered |
|---|---|---|
| Det-Seg-Refine [80] + CLIP [74] | 0.30 | 0.23 |
| GG-CNN [81] + CLIP [74] | 0.10 | 0.07 |
| GR-ConvNet [78] + CLIP [74] | 0.33 | 0.30 |
| CLIP-Fusion [53] | 0.40 | 0.40 |
| CLIPORT [18] | 0.27 | 0.30 |
| LLGD [13] | 0.43 | 0.42 |
| MaskGrasp [12] | 0.43 | 0.42 |
| GraspSAM [14] | 0.45 | 0.43 |
| **GraspMamba** (ours) | **0.54** | **0.52** |

### *E. Robotic Validation*

In Fig. 6, we present our robotic evaluation using a Kinova Gen3 7-DoF robot. The grasp detection and other techniques listed in Table III are tested using depth images from an Intel RealSense D410 depth camera. Our method predicts 4-DoF grasping poses converted to 6-DoF poses, assuming objects are placed on flat surfaces. Trajectory optimization, as outlined in [84], [85], directs the robot toward the desired poses. The inference and control processes run on an Intel Core i7 12700K processor and an NVIDIA RTX 4080S Ti graphics card. We evaluate performance in single-object and cluttered environments with various real-world objects, repeating each test 25 times for consistency across all methods. As highlighted in Table III, our approach, utilizing Mamba architecture, consistently outperforms other baselines. Notably, despite being trained exclusively on Grasp-Anything, a synthetic dataset generated using foundational models, it demonstrates strong performance on real-world objects.

### *F. Discussion*

**Limitation.** While our method demonstrates notable results regarding inference time and grasp pose accuracy in both 'Seen' and real-robot applications, it still struggles with predicting correct grasping poses for objects in complex real-world scenes, particularly within the 'Unseen' category. Although our approach shows promise in identifying the attention area on the target object, faulty grasping poses often arise when the target object is in cluttered environments

where visually similar distractors are present, as shown in Fig. 5. This indicates that the challenge lies in parsing ambiguous language and robustly aligning clear textual instructions with intricate visual features under high scene complexity. For example, given the instruction "grasp me the stapler", the model struggles to distinguish the stapler from a nearby pencil. Addressing this limitation will require enhanced feature representations and domain adaptation strategies to better generalize to novel object categories.

**Future Work.** While our study proposes a new method that integrates textual and visual features using the Mamba architecture for grasping pose detection, several promising avenues remain. First, we aim to extend our method to handle tasks in 3D space, including 3D point clouds and RGB-D images, to overcome the limitations of depth information in robotic applications. Additionally, we plan to investigate the SIM2real gap by fine-tuning our method on a smaller real-world dataset leveraging the two versions of the Grasp-Anything dataset [7] to determine if the performance gap between synthetic and real-world scenarios can be closed. Moreover, bridging the gap between the semantic concepts in text prompts and input images could enhance speed, efficiency, and hardware optimization, particularly for processing long sequences. For instance, a key goal is enabling the robot to comprehend and analyze complex human instructions and make accurate decisions quickly without relying on high-powered, energy-inefficient hardware. These approaches offer significant potential for advancing the capabilities of language-driven robotic grasping systems.

## V. Conclusion

We introduce a new vision-language model based on the MambaVision architecture for the language-driven grasp detection task. Our approach employs hierarchical feature fusion of text and image inputs, effectively integrating visual and textual information to improve grasping accuracy and inference speed. Our method achieves high precision by focusing on key regions highlighted by text guidance prompts. Extensive experiments demonstrate that our approach significantly outperforms existing baselines in vision-based benchmarks and real-world robotic grasping tests. Our code and model will be released.

## References


[1] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *ICRA*, 2021.

[2] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *ICRA*, 2022.

[3] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *ICRA*, 2022.

[4] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *ICRA*, 2015.

[5] M. Gilles, Y. Chen, E. Z. Zeng, Y. Wu, K. Furmans, A. Wong, and R. Rayyes, "Metagraspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping," *TASE*, 2023.

[6] F. Schirmer, P. Kranz, B. Bhat, C. G. Rose, J. Schmitt, and T. Kaupp, "Towards a path planning and communication framework for seamless human-robot assembly," in *HRI*, 2024.

[7] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," in *ICRA*, 2024.

[8] W. Yuan, A. Murali, A. Mousavian, and D. Fox, "M2t2: Multi-task masked transformer for object-centric pick and place," in *CoRL*, 2023.

[9] A. D. Vuong, M. N. Vu, B. Huang, N. Nguyen, H. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection," in *CVPR*, 2024.

[10] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction for robotics," *arXiv*, 2024.

[11] A. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration0," in *ICRA*, 2024.

[12] T. V. Vo, M. N. Vu, B. Huang, A. Vuong, N. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection with mask-guided attention," in *IROS*, 2024.

[13] N. Nguyen, M. N. Vu, B. Huang, A. Vuong, N. Le, T. Vo, and A. Nguyen, "Lightweight language-driven grasp detection using conditional consistency model," in *IROS*, 2024.

[14] S. Noh, J. Kim, D. Nam, S. Back, R. Kang, and K. Lee, "Graspsam: When segment anything model meets grasp detection," in *ICRA*, 2025.

[15] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *ICRA*, 2024.

[16] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *CVPR*, 2024.

[17] OpenAI, "Introducing ChatGPT," Software, accessed: February 6th 2023.

[18] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *CoRL*, 2022.

[19] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *CVPR Workshops*, 2024.

[20] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv*, 2025.

[21] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong, "A joint modeling of vision-language-action for target-oriented grasping in clutter," *arXiv*, 2023.

[22] R. Mirjalili, M. Krawez, S. Silenzi, Y. Blei, and W. Burgard, "Lan-grasp: Using large language models for semantic object grasping," *arXiv*, 2023.

[23] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *arXiv*, 2023.

[24] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv*, 2024.

[25] X. Han, Y. Tang, Z. Wang, and X. Li, "Mamba3d: Enhancing local features for 3d point cloud analysis via state space model," *arXiv*, 2024.

[26] J. Liu, M. Liu, Z. Wang, L. Lee, K. Zhou, P. An, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Multimodal state space model for efficient robot reasoning and manipulation," *arXiv*, 2024.

[27] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirom, Y. Belinkov, S. Shalev-Shwartz, O. Abend, R. Alon, T. Asida, A. Bergman, R. Glozman, M. Gokhman, A. Manevich, N. Ratner, N. Rozen, E. Shwartz, M. Zusman, and Y. Shoham, "Jamba: A hybrid transformer-mamba language model," *arXiv*, 2024.

[28] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, "Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning," *arXiv*, 2024.

[29] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "Rsmamba: Remote sensing image classification with state space model," *arXiv*, 2024.

[30] G. Wang, X. Zhang, Z. Peng, T. Zhang, and L. Jiao, "$S^2$mamba: A spatial-spectral state space model for hyperspectral image classification," *arXiv*, 2024.

[31] X. Ma, X. Zhang, and M.-O. Pun, "Rs3mamba: Visual state space model for remote sensing images semantic segmentation," *arXiv*, 2024.

[32] Q. Zhu, Y. Cai, Y. Fang, Y. Yang, C. Chen, L. Fan, and A. Nguyen, "Samba: Semantic segmentation of remotely sensed images with state space model," *arXiv*, 2024.

[33] Z. Wan, Y. Wang, S. Yong, P. Zhang, S. Stepputtis, K. Sycara, and Y. Xie, "Sigma: Siamese mamba network for multi-modal semantic segmentation," *arXiv*, 2024.

[34] D. Liang, X. Zhou, W. Xu, X. Zhu, Z. Zou, X. Ye, X. Tan, and X. Bai, "Pointmamba: A simple state space model for point cloud analysis," *arXiv*, 2024.

[35] Z. Wang, Z. Chen, Y. Wu, Z. Zhao, L. Zhou, and D. Xu, "Pointramba: A hybrid transformer-mamba framework for point cloud analysis," *arXiv*, 2024.

[36] X. Jia, Q. Wang, A. Donat, B. Xing, G. Li, H. Zhou, O. Celik, D. Blessing, R. Lioutikov, and G. Neumann, "Mail: Improving imitation learning with mamba," *arXiv*, 2024.

[37] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *CVPR*, 2019.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[39] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers," in *ICRA*, 2014.

[40] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, 2014.

[41] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *IROS*, 2018.

[42] S. Yu, D.-H. Zhai, and Y. Xia, "Egnet: Efficient robotic grasp detection network," *IEEE Transactions on Industrial Electronics*, 2023.

[43] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *ICRA*, 2017.

[44] D. Wei, J. Cao, and Y. Gu, "Robot grasp in cluttered scene using a multi-stage deep learning model," *IEEE Robotics and Automation Letters*, 2024.

[45] S. Yu, D.-H. Zhai, Y. Xia, H. Wu, and J. Liao, "Se-resunet: A novel robotic grasp detection method," *IEEE Robotics and Automation Letters*, 2022.

[46] L. Tong, K. Song, H. Tian, Y. Man, Y. Yan, and Q. Meng, "A novel rgb-d cross-background robot grasp detection dataset and background-adaptive grasping network," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[47] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *ICCV*, 2021.

[48] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds," in *ICRA*, 2020.

[49] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang, "Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes," in *IROS*, 2023.

[50] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, "Language guided robotic grasping with fine-grained instructions," in *IROS*, 2023.

[51] T. Nguyen, M. N. Vu, B. Huang, A. Vuong, Q. Vuong, N. Le, T. Vo, and A. Nguyen, "Language-driven 6-dof grasp detection using negative prompt guidance," in *ECCV*, 2024.

[52] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *ICRA*, 2022.

[53] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Wang, and R. Xiong, "A joint modeling of vision-language-action for target-oriented grasping in clutter," *arXiv*, 2023.

[54] Y. Yang, X. Lou, and C. Choi, "Interactive robotic grasping with attribute-guided disambiguation," *arXiv*, 2022.

[55] M. Zhao, G. Zuo, S. Yu, Y. Luo, C. Liu, and D. Gong, "Language-guided category push–grasp synergy learning in clutter by efficiently perceiving object manipulation space," *IEEE Transactions on Industrial Informatics*, 2025.

[56] V. Bhat, P. Krishnamurthy, R. Karri, and F. Khorrami, "Hifi-cs: Towards open vocabulary visual grounding for robotic grasping using vision-language models," *arXiv*, 2024.

[57] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal lstm," in *CVPR*, 2016.

[58] Y. Huang, Q. Wu, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *CVPR*, 2017.

[59] Y. Wan, W. Wang, G. Zou, and B. Zhang, "Cross-modal feature alignment and fusion for composed image retrieval," in *CVPR Workshops*, 2024.

[60] Y. Liu, S. Liu, B. Chen, Z.-X. Yang, and S. Xu, "Fusion-perception-to-action transformer: Enhancing robotic manipulation with 3d visual fusion attention and proprioception," *IEEE Transactions on Robotics*, 2025.

[61] X. Xu, Y. Wang, Y. He, Y. Yang, A. Hanjalic, and H. T. Shen, "Cross-modal hybrid feature fusion for image-sentence matching," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2021.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[63] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv*, 2024.

[64] C. Yang, Z. Chen, M. Espinosa, L. Ericsson, Z. Wang, J. Liu, and E. J. Crowley, "Plainmamba: Improving non-hierarchical mamba in visual recognition," *arXiv*, 2024.

[65] X. Pei, T. Huang, and C. Xu, "Efficientvmamba: Atrous selective scan for light weight visual mamba," *arXiv*, 2024.

[66] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv*, 2024.

[67] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *CVPR*, 2025.

[68] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *arXiv*, 2024.

[69] Y. Qiao, Z. Yu, L. Guo, S. Chen, Z. Zhao, M. Sun, Q. Wu, and J. Liu, "Vl-mamba: Exploring state space models for multimodal learning," *arXiv*, 2024.

[70] W. Huang, J. Pan, J. Tang, Y. Ding, Y. Xing, Y. Wang, Z. Wang, and J. Hu, "Ml-mamba: Efficient multi-modal large language model utilizing mamba-2," *arXiv*, 2024.

[71] R. Xu, S. Yang, Y. Wang, B. Du, and H. Chen, "A survey on vision mamba: Models, applications and challenges," *arXiv*, 2024.

[72] G. Zhong, W. Ding, L. Chen, Y. Wang, and Y.-F. Yu, "Multi-scale attention generative adversarial network for medical image enhancement," *TETCI*, 2023.

[73] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IROS*, 2018.

[74] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[75] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *ICCV*, 2023.

[76] Z. Li, H. Pan, K. Zhang, Y. Wang, and F. Yu, "Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion," *arXiv*, 2024.

[77] W. Li, H. Zhou, J. Yu, Z. Song, and W. Yang, "Coupled mamba: Enhanced multi-modal fusion with coupled state space model," *arXiv*, 2024.

[78] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *IROS*, 2020.

[79] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022.

[80] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *ICRA*, 2021.

[81] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv*, 2018.

[82] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *CVPR*, 2023.

[83] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[84] F. Beck, M. N. Vu, C. Hartl-Nesic, and A. Kugi, "Singularity avoidance with application to online trajectory optimization for serial manipulators," *IFAC-PapersOnLine*, 2023.

[85] M. Vu, F. Beck, M. Schwegel, C. Hartl-Nesic, A. Nguyen, and A. Kugi, "Machine learning-based framework for optimally solving the analytical inverse kinematics for redundant manipulators," *Mechatronics*, 2023.