Task-Conditioned Grasp Pose Selection Using a Structured Vision-Language Model Critic

Bo-Kai Peng
Department of Electrical Engineering
National Taiwan Normal University
Taipei, Taiwan
61375004h@ntnu.edu.tw

Chen-Chien Hsu
Department of Electrical Engineering
National Taiwan Normal University
Taipei, Taiwan
jhsu@ntnu.edu.tw

Wei-Yen Wang
Department of Electrical Engineering
National Taiwan Normal University
Taipei, Taiwan
wywang@ntnu.edu.tw

Shao-Kang Huang
Department of Artificial Intelligence
Tamkang University
Taipei, Taiwan
169394@o365.tku.edu.tw

***Abstract*—Robotic manipulation often requires grasps that are not only geometrically stable but also compatible with downstream task requirements. For example, a robot should avoid blocking a mug opening during pouring and should grasp the handle rather than the head when using a hammer. To address this problem, we propose a training-free, task-conditioned grasp selection framework based on structured vision-language reasoning. The system follows a conservative generate-then-reason design: an upstream semantic-isolation module uses the task instruction and RGB observation to constrain grasp generation to task-relevant regions, and a GraspNet-based proposer then generates 6-DoF candidates from the RGB-D observation. After reachability filtering, compact descriptors, including proposal score, approach tilt, coarse arm alignment, and semantic height cues, are provided to a structured VLM critic. The critic evaluates visual affordance and physical constraints through a strict JSON interface, enabling auditable task-conditioned decisions without directly generating continuous robot actions. We evaluate the system on a TM5-700 robot with a fixed overhead RGB-D camera across three tabletop tasks: travel mug pouring, kettle pouring, and hammer tool use. Across 120 real-robot trials with four ablation settings, the full system improves the task-valid executed-selection rate from 12/30 trials for the geometric baseline to 23/30 trials. These results suggest that semantic isolation and structured candidate auditing provide complementary benefits, while remaining failures highlight the need for stronger whole-arm feasibility modeling.**

***Keywords*: robot grasping; task-conditioned selection; vision-language model; semantic isolation; reachability filtering; structured output.**

I. INTRODUCTION

Learning-based 6-DoF grasp generation has made it practical to propose geometrically stable grasps directly from RGB-D observations. However, in real robotic manipulation, grasping is rarely the final goal. It is usually an intermediate step toward a downstream task such as pouring, tool use, or object transfer. As a result, the highest-scoring grasp under a generic geometric metric is not always the most useful grasp for the task.

This mismatch appears in two forms. First, there is a semantic gap: the same object may require different functional grasp regions depending on the instruction. For example, a hammer should be grasped by the handle for tool use, while grasping the head defeats the task purpose. Similarly, a travel mug grasp that blocks the top opening may be stable but invalid for pouring. Second, there is a physical execution gap: a grasp that appears locally valid at the gripper level can still be difficult to execute because of tabletop contact, awkward wrist posture, reachability limits, or whole-arm motion constraints.

To address these gaps, we propose a training-free, task-conditioned grasp selection framework based on structured vision-language reasoning. Instead of asking a vision-language model (VLM) to directly output a continuous 6-DoF robot pose, the system follows a generate-then-reason strategy. A grasp proposal module first generates a finite set of candidate poses, and a structured VLM critic then selects from this bounded candidate set using visual context and compact physical descriptors. This design keeps the decision process auditable and compatible with existing grasp generation and robot control pipelines.

The proposed framework further introduces an upstream semantic-isolation module before final candidate auditing. Given a task instruction and an RGB observation, this module constrains grasp generation to task-relevant object regions, such as a container body, top handle, or tool handle. The remaining candidates are then evaluated using descriptors including proposal score, approach tilt, coarse arm alignment, and semantic height cues through a strict JSON-based interface.

The main contributions of this paper are summarized as follows:

- **Two-stage VLM-mediated candidate shaping and auditing:** We formulate task-conditioned grasp selection as a two-stage decision process that combines upstream semantic candidate-pool shaping with downstream descriptor-grounded VLM auditing. This differs from directly asking a VLM to output a robot pose or merely reranking candidates after grasp generation.
- **Structured and Interpretable Candidate Auditing:** Compact physical descriptors and a strict JSON-based decision interface are introduced to support auditable grasp selection. The VLM critic evaluates candidates according to safety, task affordance, functional clearance, and geometric quality.
- **Expanded Real-Robot Ablation Study:** The framework is evaluated on a TM5-700 robot with a fixed overhead RGB-D camera across three tabletop tasks: travel mug pouring, kettle pouring, and hammer tool use. Across 120 physical trials and four ablation settings, the results show that semantic isolation and structured candidate auditing provide complementary benefits for task-valid executed grasp selection.

II. RELATED WORK

Learning-based robotic grasping has advanced from geometry-based grasp planning and visual action mapping [1]–

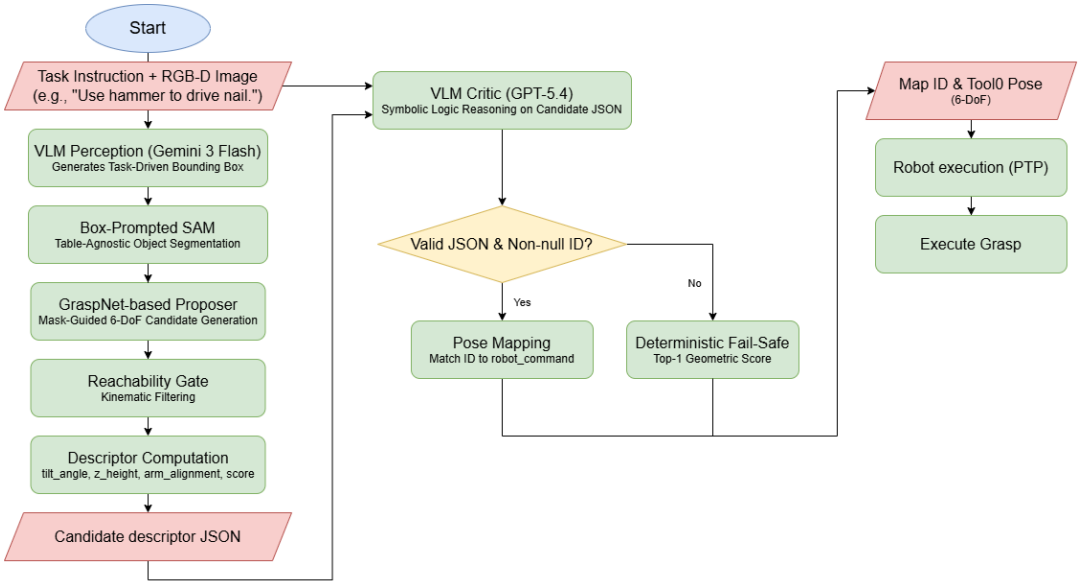

Fig. 1. Overview of the proposed generate-then-reason pipeline with semantic isolation, mask-guided grasp proposal, reachability filtering, structured VLM auditing, and conservative robot execution.

[4] to large-scale 6-DoF grasp proposal and optimization methods such as GraspNet-1Billion [5] and SE(3)-DiffusionFields [6]. These methods provide strong foundations for generating geometrically stable grasp candidates from RGB-D or point-cloud observations. However, their objectives are often centered on generic grasp stability or motion cost, rather than selecting a grasp that satisfies a downstream functional task. For example, a high-scoring grasp may still block a mug opening or grasp the wrong part of a tool.

Language grounding and promptable segmentation methods, such as SAM [7] and LSeg [8], make it possible to localize task-relevant object regions from visual or language prompts. Meanwhile, embodied VLM and VLA systems such as PaLM-E [9] and OpenVLA [10] show the potential of foundation models for robot reasoning and action generation. However, directly generating continuous robot actions often requires careful action representations, task-specific data, or additional execution constraints. In contrast, our work uses a general-purpose VLM, implemented with GPT-5.4 [11], as a structured critic over a bounded set of pre-computed grasp candidates. This training-free and discrete formulation keeps the decision process auditable while remaining compatible with existing grasp generation and robot control pipelines.

III. METHOD

The proposed pipeline maps an RGB-D observation and a natural-language instruction to a task-conditioned 6-DoF grasp candidate that can be executed by the robot. As illustrated in Fig. 1, the system separates continuous geometric grasp generation from discrete and auditable task reasoning. Instead of asking a vision-language model (VLM) to directly generate a continuous robot pose, the system first constructs a bounded candidate set and then uses structured vision-language reasoning to select a task-valid grasp ID.

The pipeline first performs task-conditioned semantic isolation to localize the target object, the grasping region, and the functional region that should be preserved or avoided. A promptable segmentation module isolates the selected region before GraspNet-based grasp proposal. The remaining candidates are then filtered by reachability, summarized with compact descriptors, grounded with numeric IDs on the RGB image, and evaluated by a structured VLM critic through a strict JSON interface. If the VLM output is invalid, a deterministic geometric fallback is used.

A. Task-Conditioned Semantic Isolation

Before grasp proposal, the system performs task-conditioned semantic isolation to improve the candidate pool. Given the instruction and RGB observation, the module identifies the target object, a task-relevant grasping region, and a functional or danger region. For example, the grasping region may correspond to a tool handle, a container body, or a stable object region, while the region to preserve may correspond to a hammer head, mug opening, or kettle spout. The selected grasping region is then used as a prompt for segmentation, producing a binary mask for downstream grasp generation, as shown in Fig. 2(a).

This stage also supports ablation. When semantic isolation is disabled, the task-relevant grasping region is set equal to the whole-object region. In this degraded setting, the proposer still receives a valid object mask, but no task-specific semantic preference is injected before grasp generation.

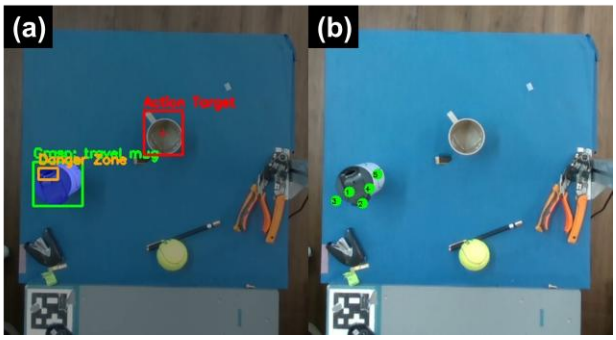

Fig. 2. Task-conditioned semantic isolation and candidate ID grounding: (a) region-level task cues; (b) ID-labeled reachable grasp candidates.

B. Grasp Proposal and Reachability Gate

After semantic isolation, a GraspNet-based proposer generates 6-DoF grasp candidates from the masked RGB-D observation. Each candidate contains a camera-frame grasp center, a rotation matrix, a projected image location, and a geometric proposal score. The grasp pose is transformed from

the camera frame to the robot base frame using the calibrated camera-to-base transformation.

To remove clearly infeasible poses before VLM reasoning, we apply a lightweight reachability gate in the robot base frame. Let the mapped translation be (t_x, t_y, t_z), and let the planar distance from the robot base be $d = \sqrt{t_x^2 + t_y^2}$.

A candidate is retained only if its planar distance and height satisfy predefined workspace limits. This gate serves as a conservative pre-filter rather than a complete motion planner.

C. Descriptor Computation and Candidate ID Grounding

For each reachable candidate, we compute a compact descriptor for VLM reasoning. Each descriptor contains the candidate ID, projected image location, proposal score, approach tilt angle, coarse arm-alignment label, physical height, and semantic height tag. These descriptors convert continuous 6-DoF grasp information into prompt-friendly physical cues.

To reduce spatial ambiguity, reachable candidates are overlaid on the RGB image using numeric IDs, as shown in Fig. 2(b). The same IDs are used in the descriptor JSON, allowing the VLM critic to connect visual candidate locations with their physical attributes. After filtering, candidates are re-indexed with consecutive integers to simplify the mapping between the image, descriptor list, and final robot command.

D. Decision Hierarchy

To guide the VLM critic toward consistent decisions, we enforce the prompted hierarchy shown in Fig. 3. The critic evaluates candidates in the following priority order:

- Level 1 (Safety): Reject unsafe or collision-prone grasps, including candidates with unfavorable approach geometry or high tabletop contact risk.
- Level 2 (Task Validity): Preserve the task-critical functional region. For example, a pouring task should keep the opening or spout clear, while a tool-use task should leave the working part unobstructed.
- Level 3 (Optimization): Among candidates that satisfy safety and task validity, use geometric cues such as proposal score, tilt angle, arm alignment, and semantic height to choose the final candidate.

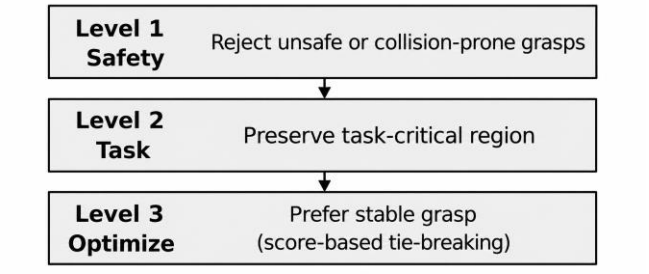

Fig. 3. Structured VLM auditing hierarchy: safety filtering, task-conditioned validity, and score-based tie-breaking.

E. VLM Critic and Output Schema

The VLM critic receives three inputs: the task instruction, the RGB image with candidate IDs, and a compact JSON list of candidate descriptors. The critic does not output a continuous robot pose. Instead, it selects from the bounded candidate set by returning either a final selected_id or an ordered list of preferred candidate IDs.

To ensure auditability, the response must be parseable JSON with explicit reasoning fields and a final candidate decision. In the updated ranked-output setting used in the candidate-level case study, the response format is defined as:

{ "functional_constraint": "...", "kinematic_analysis": "...", "ranked_ids": [<int>, ...], "reasoning": "..."}

The field functional_constraint summarizes the task requirement that must be preserved, such as keeping the mug opening unobstructed during pouring. The field kinematic_analysis explains how the candidate descriptors are interpreted, and ranked_ids provides an ordered list of candidate IDs from most to least preferred. When a ranked list is returned, the first valid ID in ranked_ids is used as the executable grasp selection.

The system validates whether the response is parseable JSON and whether the returned candidate ID or ranked IDs correspond to existing reachable candidates. Invalid, empty, null, or unparsable outputs trigger a deterministic fail-safe that selects the reachable candidate with the highest proposal score.

F. Conservative Robot Execution

After a valid candidate ID is obtained, the controller maps the ID back to the corresponding 6-DoF grasp pose T_{grasp} in the robot base frame. A pre-grasp pose T_{pre} is generated by applying a fixed retreat offset along the gripper approach direction.

The robot then executes a staged routine: move to T_{pre}, approach T_{grasp}, close the two-finger gripper, and lift the object. This conservative execution routine completes the mapping from a VLM-selected discrete ID to a physical robot action.

IV. Experiments

Hardware. Experiments were conducted on a TM5-700 robotic arm equipped with a TOYO CHY2B-S80 two-finger gripper. The workspace was observed by a fixed overhead RGB-D camera, and each trial used a single-view RGB-D observation as input. Perception modules, grasp proposal generation, and VLM inference were executed on a PC. Robot motion in the reported ablation trials was executed using a conservative staged routine based on the manufacturer-provided point-to-point (PTP) interface. The newer MoveIt-based execution extension was not included in the quantitative results reported in this paper.

Tasks. We evaluate three tabletop manipulation tasks: travel mug pouring, kettle pouring, and hammer tool use. These tasks require different forms of task-conditioned grasp reasoning. For travel mug pouring, the selected grasp should avoid blocking the top opening so that the liquid can be poured out. For kettle pouring, the grasp should preserve the pouring function while maintaining a stable and balanced grasping posture. For hammer tool use, the robot should grasp the handle and avoid grasping the hammer head, since the head must remain available for the tool-use action.

A. Ablation Protocol and Evaluation Metric

To evaluate the contribution of each module, we compare four ablation settings. Group A is the geometric baseline, where semantic isolation is degraded to a coarse whole-object region and the structured VLM critic is disabled. In this setting, the system selects the reachable candidate with the highest geometric proposal score. Group B enables semantic isolation

TABLE I

Task-wise executed-selection counts over 120 real-robot trials

Task	A Geometric	B Semantic only	C Critic only	D Full system
Travel mug pouring	3/10	3/10	4/10	7/10
Kettle pouring	5/10	5/10	5/10	8/10
Hammer tool use	4/10	7/10	6/10	8/10
Overall	12/30	15/30	15/30	23/30

but disables the VLM critic, testing whether task-aware region isolation alone improves the candidate pool. Group C disables semantic isolation but enables the structured VLM critic, testing whether downstream candidate auditing can compensate for a less task-focused candidate set. Group D is the full proposed system, combining task-conditioned semantic isolation with structured VLM-based auditing.

The experiment follows a 3 × 4 × 10 protocol: three tasks, four ablation settings, and ten physical trials for each task-setting combination, resulting in 120 real-robot trials in total. The primary metric is the task-valid executed-selection rate. A trial is counted as successful when the selected grasp can be executed by the conservative robot routine and satisfies the task requirement under post-trial inspection. This metric evaluates executable task-valid grasp selection, but not full downstream task completion such as poured liquid volume or hammer impact success. For travel mug pouring, the grasp must avoid blocking the top opening. For kettle pouring, the grasp must preserve the pouring function while maintaining a stable grasping posture. For hammer tool use, the grasp must be on the handle rather than the hammer head.

For each trial, we record the selected candidate ID, candidate descriptors, VLM response when applicable, and the final success or failure label. Results are reported as raw counts because each task-setting contains only ten trials, and raw counts avoid over-emphasizing coarse percentage values from small-sample pilot experiments.

B. Task-Wise Ablation Results and Analysis

Table I summarizes the task-wise ablation results over 120 real-robot trials. The full system, Group D, achieves the best overall performance across all three tasks, improving the task-valid executed-selection rate from 12/30 trials in the geometric baseline to 23/30 trials. This result indicates that combining upstream semantic isolation with downstream structured VLM auditing is more effective than using either component alone.

The contribution of each module differs across tasks. For travel mug pouring, the geometric baseline succeeds in 3/10 trials, while the full system succeeds in 7/10 trials. The critic-only setting also improves slightly over the baseline. This suggests that the task benefits from clearance-aware auditing, since a grasp may be geometrically stable but still invalid if it blocks the top opening. However, the largest gain appears when the candidate pool is also constrained to a task-relevant region.

For kettle pouring, Groups A, B, and C each achieve 5/10 trials, while the full system reaches 8/10 trials. This shows a complementary effect: semantic isolation alone and critic-only auditing are not sufficient, but their combination improves the result. This task requires both a reasonable candidate pool and final task-aware auditing, because a grasp must preserve the pouring function while maintaining a stable and balanced posture.

For hammer tool use, semantic isolation provides the strongest single-module improvement. The geometric baseline succeeds in 4/10 trials, while semantic isolation alone reaches 7/10 trials and the full system reaches 8/10 trials. This indicates that restricting grasp generation to the handle region directly reduces wrong-part candidates. The structured critic further improves the full system by auditing the remaining candidates, but the dominant gain comes from improving the candidate pool before final selection.

Overall, these results support the design of separating semantic isolation from candidate-level auditing. Semantic isolation controls where candidates are generated, while the structured critic controls which reachable candidate should be selected.

C. Candidate-Level Case Study

To further inspect how the structured critic makes a decision within a single scene, Table II presents a representative travel mug pouring trial under the full system. This task is useful for analysis because the highest-scoring grasp is not necessarily task-valid: a top-down grasp near the upper region of the mug may be geometrically attractive, but it can obstruct the opening required for pouring.

In this example, the geometric Top-1 candidate is ID 1, which has the highest proposal score of 0.575. However, it uses a 10° top-down approach and is located in the upper-body region. The VLM critic rejects this candidate because it may place the gripper over the top opening and block the pouring function. Instead, the critic ranks ID 2 as the first choice, followed by ID 4 and ID 3. ID 2 is a middle-body side grasp with facing-outwards alignment, which keeps the top opening clear while maintaining a feasible grasping posture.

This case study illustrates the role of the structured critic as an auditable decision layer. The final selection can be traced from the visual candidate ID to the descriptor JSON and the VLM reasoning log. More importantly, it shows that the proposed system does not simply maximize the geometric proposal score; it can reject a higher-scoring but functionally invalid grasp in favor of a task-compatible candidate.

TABLE II

Candidate-level analysis under the full system.

ID	score	tilt	alignment	z_height (cm)	z_tag	VLM decision
1	0.575	10 deg	Facing Outwards	15.2	Upper body	Rejected
2	0.364	87 deg	Facing Outwards	10.6	Middle body	Rank 1
3	0.232	92 deg	Neutral	6.6	Lower body	Rank 3
4	0.065	85 deg	Facing Outwards	11.2	Middle body	Rank 2
5	0.038	81 deg	Facing Inwards	3.6	Lower body	Rejected

D. Failure Modes and Robustness Extensions

The ablation trials reveal several remaining failure modes, which also motivate later robustness extensions in the updated implementation.

Candidate-pool and perception limitations. Some failures are caused by depth distortion, point-cloud noise, or imperfect semantic isolation. These errors can shift the generated grasp poses or leave only weak candidates after reachability filtering. Since the VLM critic selects from a bounded candidate set, it cannot recover if no task-valid and executable grasp candidate is available.

VLM judgment limitations. The structured critic may still make task-level judgment errors. For example, in container tasks, a candidate may be physically graspable but still undesirable for pouring if it blocks the functional region or leads to an unstable pouring posture. These cases show that structured VLM auditing improves task-conditioned selection, but does not eliminate all semantic or affordance-level ambiguity.

Execution feasibility limitations. Some failures are caused by execution feasibility rather than wrong task reasoning. In the ablation trials, a selected candidate can be task-valid at the grasp-selection level but still difficult to execute using the conservative PTP routine because of large wrist rotation, workspace limits, tabletop clearance, or whole-arm motion constraints. These observations motivated two later robustness extensions: allowing the critic to return a ranked list of candidate IDs instead of a single ID, and exploring MoveIt-based motion planning for execution-aware validation. However, these extensions are not included in the quantitative ablation results reported in Table I because they were developed after the 120-trial ablation study and are still being refined.

V. Discussion and Limitations

Our results highlight a pragmatic role for VLMs in manipulation: serving as an auditable critic for discrete task-conditioned grasp selection, rather than directly emitting continuous robot actions. Compared with the geometric baseline, the full system improves task-valid grasp selection across three tabletop tasks, suggesting that upstream semantic isolation and downstream structured candidate auditing provide complementary benefits. The contribution of this work is not a new grasp generator or a newly trained VLM, but a constrained and auditable interface that integrates semantic grounding, physical descriptors, and robot reachability checks for task-conditioned grasp selection.

However, the current system still has limitations. The VLM critic can only select from the generated candidate set, so it cannot recover if no task-valid candidate is proposed. Failures may also arise from perception noise, ambiguous functional-region grounding, or execution feasibility issues. These observations motivate our later ranked-candidate output and MoveIt-based execution-aware extension, which remain under refinement and are not included in the quantitative results reported in this paper.

Finally, the current evaluation focuses on internal module ablations rather than direct comparison with trained task-aware grasp rerankers. Although the structured JSON schema and decision hierarchy reduce free-form prompt sensitivity, systematic robustness tests across prompt variants, temperatures, and VLM backbones remain future work.

VI. Conclusion

We presented a training-free, task-conditioned grasp selection framework that combines semantic isolation with a structured VLM critic. Instead of directly generating continuous robot actions, the system constructs a bounded set of 6-DoF grasp candidates and selects a task-valid candidate through auditable JSON-based reasoning.

In 120 real-robot trials across travel mug pouring, kettle pouring, and hammer tool use, the full system improved the task-valid executed-selection rate from 12/30 trials for the geometric baseline to 23/30 trials. These results suggest that upstream semantic isolation and downstream structured VLM auditing are useful for task-conditioned grasp selection. Future work will focus on improving execution-aware robustness through ranked candidate retry, whole-arm motion planning, collision checking, and closed-loop visual refinement.

Acknowledgment

This work was supported in part by the “Higher Education Sprout Project” of National Taiwan Normal University and the Ministry of Education (MOE), Taiwan; and in part by the National Science and Technology Council, Taiwan, under Grant NSTC 114-2221-E-003-022-MY2 and 114-2224-E-003-001

References

[1]. H. -H. Chiang, J. -K. You, C. -C. J. Hsu and J. Jo, "Optimal Grasping Strategy for Robots With a Parallel Gripper Based on Feature Sensing of 3D Object Model," in *IEEE Access*, vol. 10, pp. 24056-24066, 2022.

[2]. Y. -W. Lin, Y. -H. Chien, M. -J. Hsu, W. -Y. Wang, C. -K. Lu and C. -C. Hsu, "Image-to-Action Translations Based Object Grasping Strategy without Depth Information and Robot Kinematics Analysis," *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, PingTung, Taiwan, 2023, pp. 53-54.

[3]. W. -Y. Wang, M. -J. Hsu, Y. -H. Chien, C. -C. Hsu, H. -H. Chiang and L. -A. Yu, "Cognitive System of a Virtual Robot Based on Perception, Memory, and Hypothesis Models for Calligraphy Writing Task," in *IEEE Access*, vol. 10, pp. 117782-117795, 2022.

[4]. P.-J. Hwang, C.-C. Hsu, P.-Y. Chou, W.-Y. Wang, and C.-H. Lin, "Vision-Based Learning from Demonstration System for Robot Arms," *Sensors*, vol. 22, 2022.

[5]. H. -S. Fang, C. Wang, M. Gou and C. Lu, "GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11441-11450.

[6]. J. Urain, N. Funk, J. Peters and G. Chalvatzaki, "SE(3)-DiffusionFields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, 2023, pp. 5923-5930.

[7]. A. Kirillov et al., "Segment Anything," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 3992-4003.

[8]. B. Li *et al.*, "Language-Driven Semantic Segmentation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[9]. D. Driess *et al.*, "PaLM-E: An Embodied Multimodal Language Model," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 8469–8488.

[10]. M. J. Kim *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[11]. OpenAI, "Introducing GPT-5.4," Mar. 2026. [Online]. Available: https://openai.com/index/introducing-gpt-5-4/ [Accessed: May 8, 2026].