Point2Act: Efficient 3D Distillation of Multimodal LLMs for Zero-Shot Context-Aware Grasping

Sang Min Kim[1] **Hyeongjun Heo**[1] **Junho Kim**[1] **Yonghyeon Lee**[2] **Young Min Kim**[1]
[1]Seoul National University, [2]Massachusetts Institute of Technology
{tkdals9082, heo0224, 82magnolia}@snu.ac.kr, yhl@mit.edu, youngmin.kim@snu.ac.kr

Abstract: We propose **Point2Act**, which directly retrieves the 3D action point relevant for a contextually described task, leveraging Multimodal Large Language Models (MLLMs). Foundation models opened the possibility for generalist robots that can perform a zero-shot task following natural language descriptions within an unseen environment. While the semantics obtained from large-scale image and language datasets provide contextual understanding in 2D images, they struggle to accurately interpret complex compositional queries and require extensive computation time. Our proposed *3D relevancy fields* bypass the high-dimensional features and instead efficiently imbue lightweight 2D point-level guidance tailored to the task-specific action. The multi-view aggregation effectively compensates for misalignments due to geometric ambiguities, such as occlusion, or semantic uncertainties inherent in the language descriptions. The output region is highly localized, reasoning fine-grained 3D spatial context that can directly transfer to an explicit position for physical action at the on-the-fly reconstruction of the scene. Our full-stack pipeline, which includes capturing, MLLM querying, 3D reconstruction, and grasp pose extraction, generates spatially grounded responses in under 20 seconds, facilitating practical manipulation tasks. Project page

1 Introduction

Robotic systems are increasingly expected to interpret and act on general, context-rich human language. Recently, the integration of vision language models (VLMs) – including CLIP [1] and Multimodal Large Language Models (MLLMs) – with 3D representations opens new possibilities for partially addressing this problem [2, 3, 4, 5, 6, 7, 8]. Despite their promise, leveraging VLMs exhibits key challenges in *efficiency* while concurrently achieving *high spatial precision*. The high dimensionality of VLM features (e.g., $>$512) renders the construction of 3D feature fields computationally expensive and memory-intensive, typically requiring 1–2 minutes per scene. Most importantly, these models often fail to interpret compositional descriptions and capture contextual nuances [5, 6].

We propose **Point2Act**, which scrutinizes the task-specific action point in 3D, conditioned on language instructions that encompass zero-shot tasks across a broad spectrum. Point2Act suggests that our ultimate objective is to accurately identify a semantically relevant 3D point for manipulation. Instead of the high-dimensional features, we prompt an MLLM to directly produce 2D-relevant points from multi-view images and efficiently aggregate them to infer 3D-relevant points. The resulting *3D relevancy field* encodes the relevancy information with precise spatial localization (see Figure 3).

To enable real-world deployment, we develop an efficient full-stack pipeline that integrates multi-view image capture, MLLM querying, 3D scene reconstruction, and grasp pose extraction. While the point-based interface is token-efficient, querying the MLLM for each view takes approximately 1–2 seconds. We address this latency by carefully pipelining the process, enabling generation of actionable 3D relevancy fields in about 20 seconds – significantly faster than comparable methods.

To summarize our contributions: (1) Point2Act distills multiview MLLM point outputs into compact 3D relevancy fields for spatial grounding robust to occlusions and view changes; (2) it supports zero-shot, context-aware tasks with part-aware, spatial, and abstract queries (e.g., "the handle of the red

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

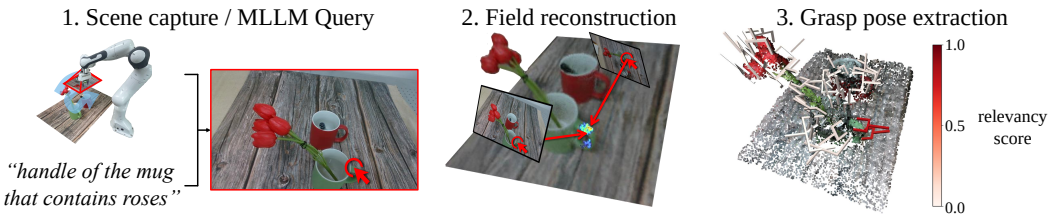


Figure 1: Overview of the **Point2Act** pipeline. We first capture posed images and query the MLLM [9] with a prompt to predict 2D point annotations on the images. The multiview predictions are distilled into a 3D relevancy field. AnyGrasp [10] proposes grasp candidates, and the most relevant grasp is selected based on the field. Grasp poses are subsampled for visualization.

mug", "the center of the monitor stand", "a dangerous part that can hurt a human hand"); and (3) it forms a practical system that runs end-to-end in ∼20s, deployable in real-world settings.

2 Method

Point2Act outputs highly localized 3D positions by combining scene context with instruction semantics, enabling context-aware grasping—predicting the appropriate 6-DoF grasp pose to satisfy a given description. We first present the 3D relevancy field for scene and action-point reconstruction (Section 2.1), then show how it guides grasp generation (Section 2.2), and finally describe the efficient, real-world-ready pipeline (Section 2.3). An overview is shown in Figure 1.

2.1 Relevancy Field Distillation from MLLM

We first convert 2D relevancy points from MLLM (e.g., Molmo [9]) into soft masks using a Gaussian kernel. Using multiview images and relevancy masks, we build a 3D relevancy field, encoding scene geometry and language-grounded relevancy. Based on Neural Radiance Fields [11] (NeRF), we add a lightweight MLP to predict relevancy values. The geometry branch is trained on RGB images to reconstruct scene structure, while the relevancy branch learns from soft masks capturing linguistic relevance. By aggregating masks from multiple views, the model overcomes occlusions and view-dependency, producing a precise, view-invariant 3D representation.

2.2 Grasping with Relevancy Fields

After learning the 3D field, we extract action poses corresponding to the instruction by relying on low-level geometric guidance rather than complex reasoning. We first convert the learned field into an RGB point cloud by rendering RGB, depth, and relevancy maps from multiple views and unprojecting them to 3D. This point cloud is fed into AnyGrasp [10] to generate grasp candidates. To select the most relevant grasp, we filter candidates by retrieving the $k = 30$ nearest neighbor points around each grasp center and choose the pose with the most relevant neighbor. This ensures the grasp is physically feasible and semantically aligned with the instruction.

2.3 Efficient System Design

We design a pipelined system (Figure 2) that captures multi-view images with a wrist-mounted camera, queries an MLLM for language-grounded points, and concurrently loads NeRF and AnyGrasp [10] to avoid delays. The 3D field is trained for 300 iterations, with grasp pose extraction starting at iteration 200 to overlap processing. Leveraging lightweight scalar relevancy supervision, the system converges quickly (Figure 4(a)) and completes the full pipeline in 16.5 seconds.

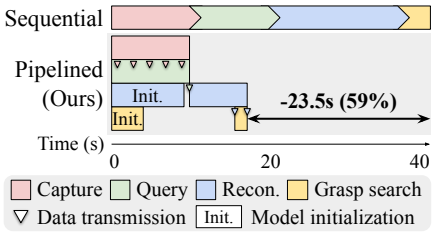


Figure 2: System diagram of Point2Act.

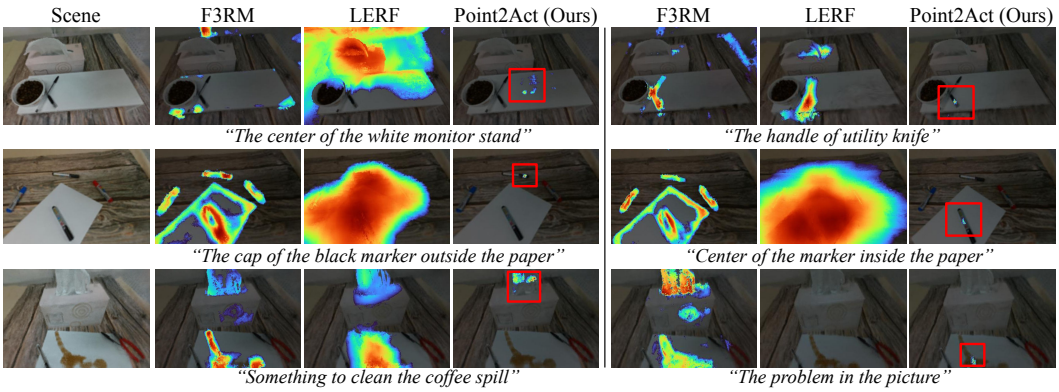
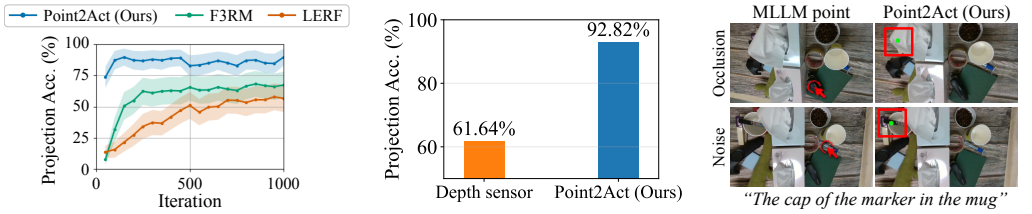


Figure 3: Comparison of language grounding methods. The first column shows the RGB image; others show relevancy scores overlaid (red: high, blue: low). Prompts are below.



(a) Effect of distillation source (b) Advantage of 3D distillation (c) Occlusion and noise reduction

Figure 4: Quantitative results for 3D localization. (a) Localization accuracy of different language grounding methods. (b) Effectiveness of our multi-view 3D distillation. (c) With multiview integration, Point2Act produces robust, view-invariant relevant 3D points from noisy MLLM predictions.

3 Experiments

We demonstrate the performance of Point2Act on a wide range of real-world examples for zero-shot context-aware grasping. We first evaluate the accuracy and efficacy of our relevancy fields (Section 3.1) and then discuss the resulting zero-shot grasping performance (Section 3.2). We then illustrate further example scenarios that require combining several contextual reasoning (Section 3.3).

3.1 Context-Aware 3D Localization

Advantages of Using MLLM Points for 3D Distillation. We evaluate the use of MLLM-predicted points as distillation targets for 3D localization, comparing against LERF [3] (multi-scale CLIP features) and F3RM [5] (MaskCLIP [12] embeddings). After training, each method selects the most relevant 3D point in the scene. Projection accuracy measures whether this point projects inside the ground-truth mask across views. As shown in Figures 3 and 4(a), Point2Act captures spatial cues, relational descriptions, and subtle intent, achieving higher projection accuracy than baselines with only **50** iterations. This demonstrates fast, precise 3D grounding from sparse, localized signals.

Effectiveness of Multi-view 3D Distillation. Point2Act benefits from multiview information not only for geometry reconstruction but also for 3D language grounding. To validate this, we compare Point2Act with a single-view MLLM+Depth baseline. Figures 4(b) and 4(c) illustrate the robustness of our approach to occlusion and viewpoint variation, even in complex real-world scenes.

3.2 Context-Aware Zero-Shot Robotic Grasping

Leveraging strong localization, Point2Act generates plausible zero-shot grasp poses without task-specific training. We compare against F3RM, LERF-TOGO, and MLLM+Depth baselines. F3RM distills CLIP features into 3D fields, while LERF-TOGO parses complex queries into object–part

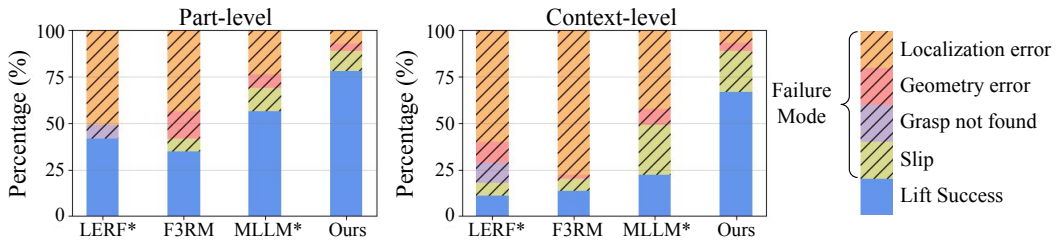
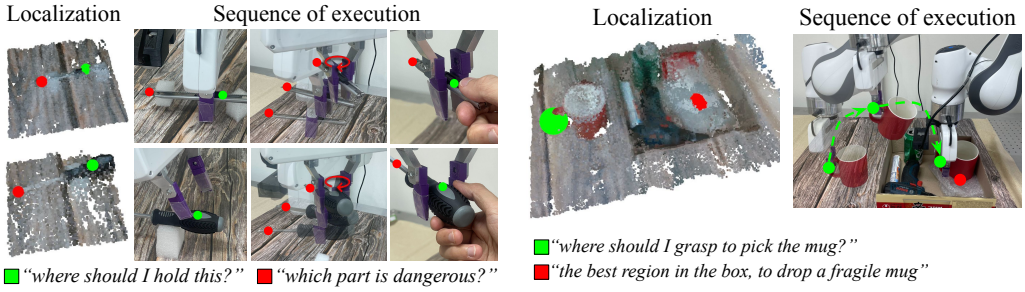


Figure 5: Grasping performance overview. Shaded areas indicate failure modes. Grasping performance evaluation is evaluated for part-level prompts (left) and context-level prompts (right). LERF* refers to LERF-TOGO [6] and MLLM* refers to MLLM 2D points with depth unprojection.



(a) Tool-Agnostic Safe Handover (b) Context-Aware Pick and Place

Figure 6: Qualitative results of (a) tool-agnostic safe handover and (b) context-aware pick-and-place. In (a), green shows graspable regions and red marks dangerous parts to avoid. In (b), the system finds grasp and safe placement areas based on context.

pairs; both lack fine-grained spatial and contextual reasoning. MLLM+Depth uses single-view predictions, making it vulnerable to occlusions. We evaluate on four real-world scenes with 20 prompts, grouped into **part-level** (grasping a specific part, e.g., "handle of the mug") and **context-level** (interpreting scene context, e.g., "the object that red scissors are pointing at", or applying commonsense, e.g., "something to clean the spilled coffee"). The results are shown in Figure 5.

3.3 Downstream Applications

We extend the relevancy field to two channels, allowing Point2Act to handle multiple prompts simultaneously. For tool-agnostic safe handover (Fig.6(a)), given prompts such as "Where should I hold this?" and "Which part is dangerous?", Point2Act identifies safe grasp points and hazardous regions to orient tools accordingly. Unlike prior object-specific methods [13, 14, 15], it generalizes to unseen tools (e.g., utility knife, screwdriver) without additional training. In context-aware pick and place (Fig. 6(b)), Point2Act locates graspable regions and safe placement areas from separate prompts, enabling flexible, context-driven manipulation without task-specific tuning.

4 Conclusion

We present Point2Act, a practical and efficient system that combines Multimodal Large Language Models (MLLMs) with 3D field representations to address the problem of zero-shot, context-aware grasping. By using point responses as communication media with MLLMs, we distill them into a highly localized and view-independent 3D relevancy field. This sparse point signal is remarkably easy to learn, significantly reducing the number of iterations required for 3D field reconstruction. We demonstrate that the resulting 3D relevancy field enables precise spatial localization, providing a reliable guidance signal for zero-shot grasping. With our pipelined system design, the full pipeline runs in under 20 seconds, highlighting the practicality of our approach. Finally, we show that Point2Act can be extended to multi-channel settings, enabling downstream applications such as tool-agnostic safe handovers and context-aware pick-and-place.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[2] S. Kobayashi, E. Matsumoto, and V. Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022.

[3] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.

[4] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[5] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola. Distilled feature fields enable few-shot language-guided manipulation. In *CoRL*, 2023.

[6] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023.

[7] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

[8] S. Koch, J. Wald, M. Colosi, N. Vaskevicius, P. Hermosilla, F. Tombari, and T. Ropinski. Relationfield: Relate anything in radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[9] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[10] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[12] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

[13] J. H. Kang, P. Limcaoco, N. Dhanaraj, and S. K. Gupta. Safe robot to human tool handover to support effective collaboration. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87363, page V008T08A088. American Society of Mechanical Engineers, 2023.

[14] Z. Wang, Z. Liu, N. Ouporov, and S. Song. Contacthandover: Contact-guided robot-to-human object handover. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9916–9923. IEEE, 2024.

[15] C. L. Blengini, P. D. C. Cheng, and M. Indri. Safe robot affordance-based grasping and handover for human-robot assistive applications. In *IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2024.