Hyperparameter-Tuned Object Grounding and Early-Stopped Motion Control Enhance 3D-LOTUS++ on GemBench

CVPR 2025 Gembench Challenge report

Team: MiRA
Member: Poshenc Chen, Sin-Yi Chiu, Chuan-Yu Wu, Jia-Fong Yeh,
Hung-Ting Su, Chih-Han Chen, Yan-Xiang Qiu
Contact email: poshengchen@cmlab.csie.ntu.edu.tw

Abstract

This report presents our solution to the GemBench Challenge at CVPR 2025. Based on 3D-LOTUS++, we modified the inference hyperparameters of the VLMs and applied early stopping during the training of the motion control policy. On the private test set, our method outperformed the best baseline by nearly 5%, showing significant gains in L2 (66.0) and L3 (49.4).

1. Introduction

In this challenge, GemBench offers a comprehensive suite of training and evaluation tasks designed to assess a model's generalization ability across diverse manipulation scenarios. The training set comprises 16 distinct tasks and 31 task variants, encompassing seven fundamental manipulation primitives. The test set includes 44 tasks with a total of 92 variations, categorized into four difficulty levels: novel placement, novel rigid objects, novel articulated objects, and long-horizon tasks. A portion of the test set is kept private and used for final evaluation on a hidden dataset.

Our submission is based on 3DLOTUS++. In 3DLOTUS++, we observed that the object grounding inference occasionally filtered out the target object. To address this issue, we adjusted the relevant hyperparameters. Additionally, we identified an overfitting problem during the training of the motion control policy, which we mitigated by applying early stopping.

On the GemBench public benchmark, our method achieved the following performance scores across the four levels: L1: 68.32 ± 1.04, L2: 63.57 ± 2.30, L3: 41.95 ± 1.18, and L4: 15.25 ± 1.17. On the private dataset, our method attained L2: 66.0, L3: 49.4, and L4: 15.0.

1.1. Related Work

3DLOTUS++ [2] framework:

Task Planning We define six action primitives for object manipulation. Human instructions are decomposed into a sequence of action primitives using the LLaMA-3 8B [1] LLM. By providing a few in-context examples, the LLM is guided to generate appropriate action sequences.

Object Grounding The OWLv2 [4] open-vocabulary detector generates high-confidence bounding boxes and semantic embeddings from images. These boxes are then segmented using SAM[3], and combined with RGB-D data to produce 3D point clouds. The point cloud is labeled into four categories: goal object, goal target, robot, and obstacle. Among them, the object and target are selected based on OWLv2's semantic embeddings to match the textual descriptions.

Motion Control The point cloud results from object grounding, along with the action primitives from task planning, are fed into a cross-attention encoder. This allows the model to make distinct predictions based on different targets and different actions.

2. Method

Object Grounding To refine the object grounding process during inference, we adjust several hyperparameters that control candidate box selection. The default parameters are defined as follows:

- threshold: Objectness score threshold. Only predictions with confidence scores higher than this value are retained.
- min_size_ratio: Minimum size ratio relative to the entire scene or input space. Objects smaller than this ratio are discarded to reduce noise.
- max_size_ratio: Maximum size ratio. Predictions exceeding this ratio are considered too large and removed.
- min_return_topk: Ensures that at least this number of object candidates are returned, regardless of filtering results.
- max_return_topk: Caps the number of returned object candidates to this value, preventing excessive

Table 1. Average performance (%) across GemBench Levels 2–4.

Method	L2	L3	L4
3D-LOTUS (baseline)	49.9±2.2	38.1±1.1	0.3±0.3
3D-LOTUS++ (baseline)	64.5±0.9	41.5±1.8	**17.4±0.4**
MiRA (seed200 & 300)	**66.0±0.1**	**42.7±1.3**	16.3±1.3
MiRA (ours)	63.6±2.3	**42.0±1.2**	15.3±1.2

Table 2. Performance (%) across GemBench private dataset

Method	L2	L3	L4	avg.
3D-LOTUS (baseline)	13.0	**52.2**	0.0	21.7
3D-LOTUS++ (baseline)	58.0	41.1	**17.2**	38.8
MiRA (ours)	**66.0**	49.4	15.0	**43.5**

proposals.

- use_nms: Boolean flag indicating whether to apply Non-Maximum Suppression (NMS) to reduce overlapping candidates.
- nms_sigma: Gaussian parameter for soft-NMS or suppression radius for standard NMS.
- nms_thresh: IoU threshold used during NMS to suppress overlapping boxes.

During evaluation, we observed that several target objects were mistakenly filtered out due to overly restrictive parameter settings in tasks objects (e.g., the cupboard in the Put In Cupboard task). So, we adjusted **max_return_topk** and **min_size_ratio** to mitigate this issue.

Motion Control During Motion Control training, five loss terms are tracked: total, pos, rot, open, and stop. Each converges at a different pace, resulting in distinct optimal training lengths for the respective components. We observed that the key to achieving the best performance is allowing the pos loss to converge without overfitting. Therefore, our early-stopping criterion is defined solely by the behaviour of this pos loss.

3. Experimental and Result

3.1. Experimental Setup

Object Grounding Our experiments use the set of hyperparameters:

- We keep proposals whose objectness score exceeds 0.1 (threshold = 0.1).
- Box size filtering removes extremely small and large candidates: we discard any box whose projected area is < 0.15 % of the scene (min_size_ratio = 0.0015) or > 80 % (max_size_ratio = 0.8).
- To guarantee coverage, we always return at least one proposal but never more than twenty (min_return_topk = 1, max_return_topk = 20).
- Non-maximum suppression is enabled (use_nms = True) with a Gaussian soft-NMS decay of $\sigma = 0.2$ and an IoU suppression threshold of 0.1 (nms_sigma = 0.2, nms_thresh = 0.1).

The hyperparameters that differ from the 3D-LOTUS++ setup are max_return_topk and min_size_ratio. We observed that the original settings also caused some intended targets to be removed (for example, the cupboard in the "Put In Cupboard" task), which led to incorrect point-cloud labeling and ultimately affected action prediction. To remedy this, we increased max_return_topk from 10 to 20. This change, however, introduced many small spurious boxes, so we also raised min_size_ratio from 0 to 0.0015.

Motion Control From the training loss curves, we noticed that the validation rotation loss began overfitting very early, model_step_30000 already gave the lowest value. The position loss converged at roughly 127000 steps and started overfitting by 140000 steps (the baseline checkpoint). After evaluating model_step_30000 and model_step_127000, we chose the better-performing model_step_127000 as our submission model.

3.2. result

We evaluated our method on the GemBench public test set. As shown in Table 1, our approach demonstrates a slight decrease in the overall average task success rate across all seeds compared to the official baseline. However, our method exhibits modest improvements specifically on seed 200 & 300, suggesting potential under certain initialization conditions. The detailed results for each task are provided in the Supplementary Material.

As shown in Table 2, our performance on the private test dataset improved by nearly 5% compared to the best-performing baseline. Specifically, when compared to the top baseline 3D-LOTUS++, our method achieved noticeable gains in both L2 and L3 tasks. However, performance on L4 still lags slightly behind.

4. Conclusion

Based on 3D-LOTUS++, we tuned its hyperparameters and achieved better performance than the official baseline on the private dataset. We identified shortcomings in both the object-grounding and motion-control modules and introduced improvements. Although additional experiments are required to verify that our adjustments are optimal, we believe these two components still offer considerable room for further research toward more broadly generalized robotics.

References

[1] AI@Meta. Llama 3 model card. 2024. 1

[2] Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In *International Conference on Robotics and Automation (ICRA)*, 2025. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[4] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36: 72983–73007, 2023. 1

Hyperparameter-Tuned Object Grounding and Early-Stopped Motion Control Enhance 3D-LOTUS++ on GemBench

Supplementary Material

Table 3. Performance on GemBench Level 1.

Method	Avg.	Close Fridge+0	Close Jar+15	Close Jar+16	CloseLaptop Lid+0	Close Microwave+0	LightBulb In+17	LightBulb In+19	Open Box+0	Open Door+0	Open Drawer+0
3D-LOTUS (baseline)	**94.3**±3.5	96±3.7	**100**±0.0	**100**±0.0	98±2.5	98±4.0	84±7.4	85±9.5	**99**±2.0	77±2.5	83±18.7
3D-LOTUS++ (baseline)	68.7±0.6	95±0.0	**100**±0.0	99±2.0	28±2.5	87±5.1	55±10.5	45±8.9	55±8.9	79±9.7	68±11.2
MiRA (ours)	68.32±6.01	95±3.16	98±2.45	99±2	84±2	84±7.35	42±5.1	48±12.08	66±10.68	79±6.63	52±11.22
Method	**Open Drawer+2**	**Pick& Lift+0**	**Pick& Lift+2**	**Pick& Lift+7**	**PickUp Cup+8**	**PickUp Cup+9**	**PickUp Cup+11**	**Push Button+0**	**Push Button+3**	**Push Button+4**	**PutIn Cupboard+0**
3D-LOTUS (baseline)	93±6.0	99±2.0	**100**±0.0	**100**±0.0	93±4.0	94±3.7	94±4.9	99±2.0	100±0.0	100±0.0	89±5.8
3D-LOTUS++ (baseline)	75±4.5	97±6.0	94±3.7	93±5.1	88±6.6	88±6.6	91±4.9	100±0.0	100±0.0	100±0.0	1±2.0
MiRA (ours)	55±13.04	98±2.45	93±6.78	90±4.47	91±3.74	84±7.35	88±6	**100**±0	**100**±0	**100**±0	9±5.83
Method	**PutIn Cupboard+3**	**PutMoney InSafe+0**	**PutMoney InSafe+1**	**Reach& Drag+14**	**Reach& Drag+18**	**Slide Block+0**	**Slide Block+1**	**Stack Blocks+30**	**Stack Blocks+36**	**Stack Blocks+39**	
3D-LOTUS (baseline)	72±11.2	94±3.7	94±3.7	100±0.0	100±0.0	100±0.0	100±0.0	91±6.6	90±4.5	90±5.8	
3D-LOTUS++ (baseline)	2±2.5	22±6.8	16±4.9	94±3.7	62±8.7	100±0.0	65±5.5	86±5.8	20±4.5	28±13.6	
MiRA (ours)	15±10.49	22±9.8	4±3.74	56±8.6	27±2.45	**100**±0	73±8.12	65±8.37	59±11.58	42±10.77	

Table 4. Performance on GemBench Level 2.

Method	Avg.	Push Button+13	Push Button+15	Push Button+17	Pick& Lift+14	Pick& Lift+16	Pick& Lift+18	PickUp Cup+10	PickUp Cup+12	PickUp Cup+13
3D-LOTUS (baseline)	49.9±2.2	99±2.0	100±0.0	100±0.0	3±2.5	18±8.7	33±9.3	89±3.7	78±8.7	57±7.5
3D-LOTUS++ (baseline)	64.5±0.9	99±2.0	100±0.0	99±2.0	94±3.7	96±3.7	95±3.2	79±4.9	89±9.7	84±10.2
MiRA (ours)	63.57±6.33	**100**±0	**100**±0	**100**±0	89±10.68	91±3.74	86±8.6	78±9.27	86±10.68	82±10.3
Method	**Stack Blocks+24**	**Stack Blocks+27**	**Stack Blocks+33**	**Slide Block+2**	**Slide Block+3**	**Close Jar+3**	**Close Jar+4**	**LightBulb In+1**	**LightBulb In+2**	**Lamp On+0**
3D-LOTUS (baseline)	13±8.1	40±9.5	69±5.8	1±2.2	1±2.2	71±5.8	90±4.5	24±4.9	81±6.6	0±0.0
3D-LOTUS++ (baseline)	22±9.3	83±7.5	63±7.3	27±9.8	5±3.2	98±2.5	98±2.5	56±9.7	43±7.5	2±2.0
MiRA (ours)	46±5.83	74±7.35	68±6.78	45±9.49	16±6.63	97±4	86±7.35	46±11.58	63±6.78	1±2
Method	**Reach& Drag+5**	**Reach& Drag+7**	**PutCube InSafe+0**	**Pick&Lift Cylinder+0**	**Pick&Lift Star+0**	**Pick&Lift Moon+0**	**Pick&Lift Toy+0**	**PutIn Cupboard+7**	**PutIn Cupboard+8**	
3D-LOTUS (baseline)	95±4.5	18±10.8	25±5.5	69±6.6	93±6.0	80±4.2	66±3.7	0±0.0	0±0.0	
3D-LOTUS++ (baseline)	94±2.0	64±12.4	37±5.1	91±2.0	94±3.7	29±6.6	71±2.0	0±0.0	0±0.0	
MiRA (ours)	42±17.2	45±0	28±11.66	88±9.27	99±2	43±5.1	75±7.07	6±3.74	0±0	

Table 5. Performance on GemBench Level 3.

Method	Avg.	Close Door+0	Close Box+0	Close Fridge2+0	CloseLaptop Lid2+0	Close Microwave2+0	Open Door2+0	Open Box2+0
3D-LOTUS (baseline)	38.1±1.1	0±0.0	58±8.1	36±9.7	54±10.7	85±7.1	42±6.8	11±6.6
3D-LOTUS++ (baseline)	41.5±1.8	1±2.0	29±8.6	93±2.5	50±9.5	99±2.0	52±10.3	16±8.0
MiRA (ours)	41.95±6.07	1±2	9±3.74	92±4	40±10	99±2	57±8.12	23±12.08

Method	Open Drawer2+0	Open Drawer3+0	OpenDrawer Long+0	OpenDrawer Long+1	OpenDrawer Long+2	OpenDrawer Long+3	Toilet SeatUp+0	Open Fridge+0
3D-LOTUS (baseline)	90±3.2	22±8.1	56±13.9	33±11.2	17±8.1	75±6.3	0±0.0	4±5.8
3D-LOTUS++ (baseline)	70±5.5	41±4.9	72±4.0	52±10.8	23±8.1	78±5.1	8±5.1	0±0.0
MiRA (ours)	71±4.9	19±5.83	81±6.63	61±4.9	37±8.12	69±5.83	12±9.8	0±0

Method	OpenLaptop Lid+0	Open Microwave+0	PutMoney InSafe+2	Open Drawer+1	Close Drawer+0	Close Grill+0
3D-LOTUS (baseline)	100±0.0	0±0.0	0±0.0	0±0.0	87±8.1	29±6.6
3D-LOTUS++ (baseline)	86±6.6	0±0.0	13±8.1	0±0.0	69±5.8	19±13.9
MiRA (ours)	75±10	0±0	22±7.48	0±0	72±7.48	41±14.63

Table 6. Performance on GemBench Level 4.

Method	Avg.	Push Buttons4+1	Push Buttons4+2	Push Buttons4+3	TakeShoes OutOfBox+0	PutItems InDrawer+0	PutItems InDrawer+2
3D-LOTUS (baseline)	0.3±0.3	3±4.0	0±0.0	0±0.0	0±0.0	0±0.0	0±0.0
3D-LOTUS++ (baseline)	17.4±0.4	76±7.4	49±8.6	37±8.1	0±0.0	0±0.0	0±0.0
MiRA (ours)	15.25±3.56	67±8.12	21±6.63	34±5.83	0±0	0±0	0±0

Method	PutItems InDrawer+4	Tower4+1	Tower4+3	Stack Cups+0	Stack Cups+3	PutAllGroceries InCupboard+0
3D-LOTUS (baseline)	0±0.0	0±0.0	0±0.0	0±0.0	0±0.0	0±0.0
3D-LOTUS++ (baseline)	0±0.0	17±10.8	30±13.4	0±0.0	0±0.0	0±0.0
MiRA (ours)	0±0	23±6.78	38±15.36	0±0	0±0	0±0