# ET-PLAN-BENCH: EMBODIED TASK-LEVEL PLANNING BENCHMARK TOWARDS SPATIAL-TEMPORAL COGNITION WITH FOUNDATION MODELS

Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, Tongtong Cao, Xingyue Quan, Yuzheng Zhuang, Yingxue Zhang, Jianye Hao

HUAWEI

# Main Contribution

- 1. Design complex planning tasks with spatio-temporal constraints and introduce new, more challenging embodied planning benchmarks.
- 2. Propose an automated generation of planning tasks with spatio-temporal constraints and a framework for automatically evaluating the performance of foundational models.
- 3. Conduct benchmarking based on various foundational models and study the effectiveness of supervised fine-tuning.

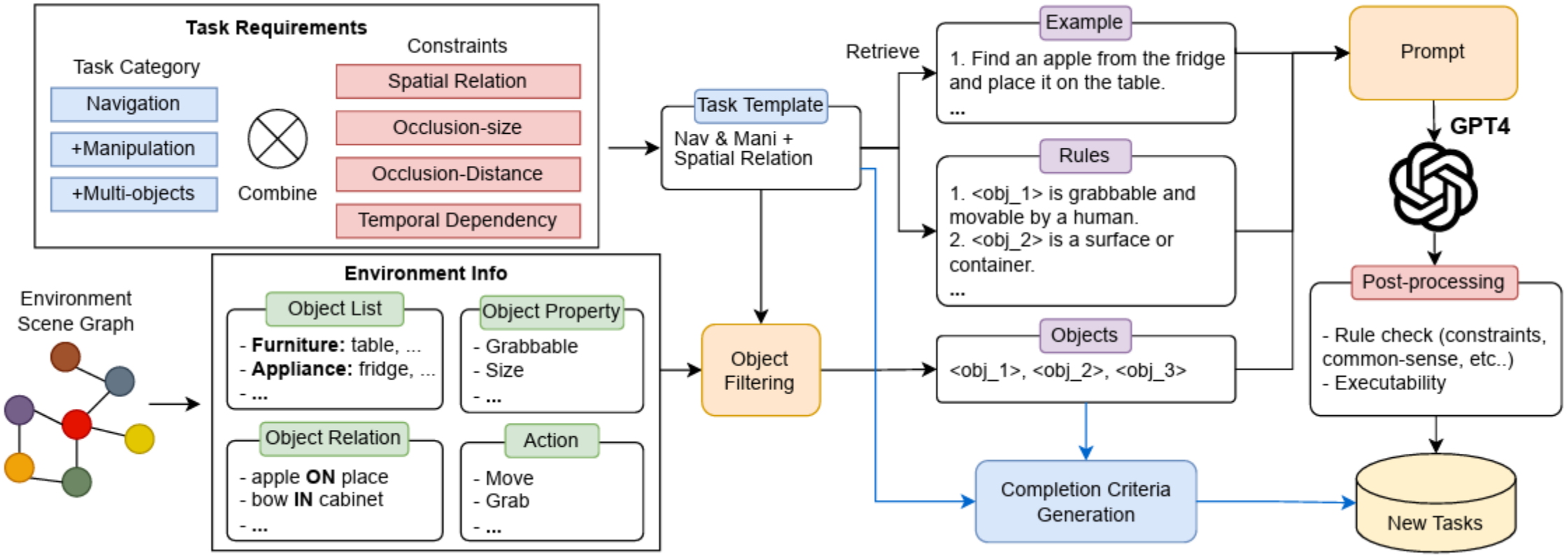
# Related Work

| Dataset | Task | Multi-Modality | Data Size | Auto Data | LLM Eval | Open Voc | Level of Obs | Spatial | Temporal/ Causal | Env Inter |
|---|---|---|---|---|---|---|---|---|---|---|
| ActivityPrograms(Puig et al., 2018) | | ✔ | 2821 | ✗ | ✗ | ✔ | P | ✗ | ✗ | ✔ |
| WAH(Puig et al., 2020) | | ✗ | 1211 | ✗ | ✗ | ✗ | P/ G | ✗ | ✗ | ✔ |
| ALFRED(Shridhar et al., 2020a) | | ✔ | 8055 | ✗ | ✗ | ✔ | P | ✗ | ✗ | ✔ |
| WAH-NL(Choi et al., 2024) | | ✗ | 611 | ✗ | ✔ | ✔ | P | ✗ | ✗ | ✔ |
| RoboGen(Wang et al., 2023) | | ✔ | ∞ | ✔ | ✔ | ✔ | P | ✗ | ✗ | ✔ |
| BEHAVIOR(Srivastava et al., 2022) | | ✗ | 100 | ✗ | ✗ | ✗ | P | ✗ | ✗ | ✔ |
| Mini-BEHAVIOR(Jin et al., 2023) | TGP | ✗ | 20 | ✗ | ✗ | ✗ | P | ✗ | ✗ | ✔ |
| BEHAVIOR-1K(Li et al., 2023) | | ✗ | 1000 | ✗ | ✗ | ✗ | P | ✗ | ✗ | ✔ |
| EgoCOT(Mu et al., 2024) | | ✔ | 129 | ✗ | ✔ | ✔ | P | ✗ | ✗ | ✗ |
| EgoPlan-Bench(Chen et al., 2023) | | ✔ | 2406 | ✔ | ✔ | ✔ | P | ✗ | ✗ | ✔ |
| EgoPlan-IT(Chen et al., 2023) | | ✔ | 50K | ✔ | ✔ | ✔ | P | ✗ | ✗ | ✔ |
| HandMeThat(Wan et al., 2022) | | ✗ | 300K | ✔ | ✗ | ✔ | P/ G | ✗ | ✗ | ✔ |
| EgoVQA(Fan, 2019) | | ✔ | 520 | ✗ | ✗ | ✔ | P/ G | ✗ | ✗ | ✗ |
| EgoTaskQA(Jia et al., 2022) | EQA | ✔ | 40K | ✗ | ✗ | ✔ | P | ✔ | ✔ | ✗ |
| Egothink(Cheng et al., 2024) | | ✔ | 700 | ✗ | ✔ | ✔ | P | ✔ | ✔ | ✗ |
| OpenEQA(Majumdar et al., 2024) | | ✔ | 1600 | ✗ | ✔ | ✔ | P | ✗ | ✗ | ✔ |
| **ET-Plan-Bench** | TGP | ✔ | ∞ | ✔ | ✔ | ✔ | P/ FP/ G | ✔ | ✔ | ✔ |

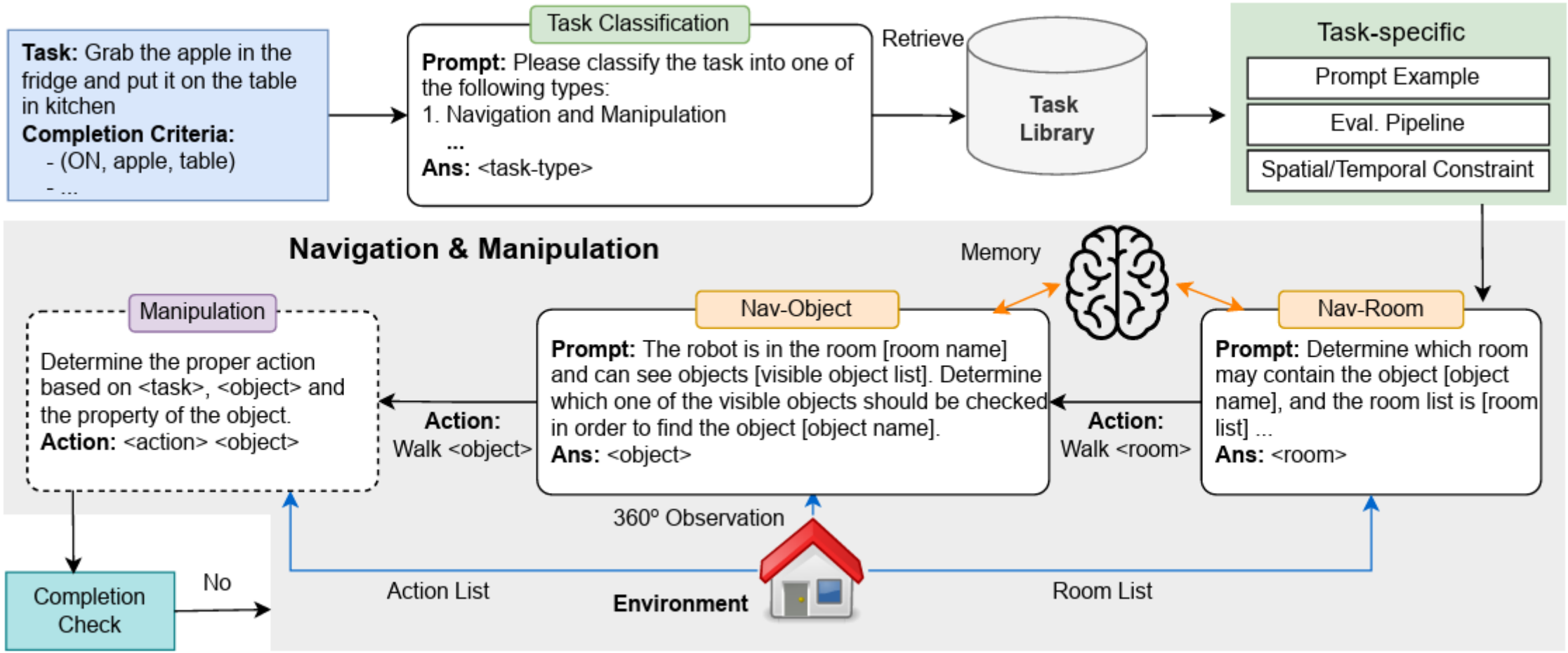
# Task Difficulty Dimension

- Action Sequence Length
- Prior Knowledge
- Spatial Relationship Constraints
- Occlusion
- Temporal and Causal Relationship Constraints

# Embodied Task Generation Pipeline



# LLM Agent Pipeline



# Experiment Results

| | Success Rate | | Seq Length | | Longest Common Seq (Ratio) | | Moving Distance | |
|---|---|---|---|---|---|---|---|---|
| Tasks | GPT4 | LLAMA 7B SFT | GPT4 | LLAMA 7B SFT | GPT4 | LLAMA 7B + SFT | GPT4 | LLAMA 7B+ SFT |
| Navigation Tasks with or without Spatial Constraints | | | | | | | | |
| Navi + *Layout Map* | 90.77% | 91.13% | 3.76 | 3.96 | 1.32 (89.64%) | 1.34 (90.11%) | 10.79 | 10.83 |
| Navi | 79.26% | 80.58% | 6.77 | 6.75 | 1.59 (78.74%) | 1.62 (80.01%) | 14.10 | 14.59 |
| Navi + *Occlusion_Size* | 72.46% | 76.05% | 7.99 | 7.89 | 1.53 (74.95%) | 1.59 (78.14%) | 14.08 | 16.65 |
| Navi + *Occlusion_Distance* | 73.65% | 76.65% | 7.94 | 7.69 | 1.60 (77.40%) | 1.65 (80.24%) | 19.36 | 17.38 |
| Navi + *Relation* | 62.61% | 64.09% | 9.20 | 9.20 | 1.78 (88.75%) | 1.75 (86.05%) | 12.45 | 14.08 |
| Navi + *Relation* + *Occlusion_Size* | 60.74% | 61.48% | 9.68 | 9.91 | 1.74 (85.19%) | 1.70 (83.21%) | 13.26 | 16.05 |
| Navi + *Relation* + *Occlusion_Distance* | 54.81% | 55.56% | 10.41 | 10.31 | 1.73 (86.67%) | 1.67 (82.96%) | 15.75 | 16.22 |
| Navigation & Manipulation Tasks with or without Spatial Constraints | | | | | | | | |
| Navi & Mani + *Layout Map* | 83.98% | 83.96% | 12.36 | 12.09 | 4.20 (82.68%) | 4.17 (81.97%) | 22.22 | 21.67 |
| Navi & Mani | 73.76% | 74.33% | 17.02 | 16.47 | 4.17 (78.56%) | 4.22 (78.92%) | 28.51 | 26.99 |
| Navi & Mani + *Occlusion_Size* | 65.85% | 67.60% | 20.00 | 19.21 | 3.94 (75.00%) | 4.00 (75.27%) | 29.46 | 28.79 |
| Navi & Mani + *Occlusion_Distance* | 72.09% | 74.66% | 18.83 | 17.20 | 4.06 (75.50%) | 4.21 (78.08%) | 37.92 | 30.45 |
| Navi & Mani + *Relation* | 49.65% | 50.35% | 24.08 | 23.78 | 4.11 (73.60%) | 4.12 (72.14%) | 27.70 | 27.72 |
| Navi & Mani + *Relation* + *Occlusion_Size* | 43.03% | 42.75% | 26.52 | 26.39 | 3.81 (69.02%) | 3.82 (67.05%) | 31.55 | 28.42 |
| Navi & Mani + *Relation* + *Occlusion_Distance* | 49.88% | 50.00% | 23.96 | 23.79 | 4.20 (73.69%) | 4.25 (72.57%) | 34.85 | 30.83 |
| Navigation & Manipulation Tasks with Temporal Constraints | | | | | | | | |
| Navi & Mani + *Multi Objects* | 56.73% | 55.05% | 38.25 | 42.80 | 7.57 (69.92%) | 7.85 (69.09%) | 50.39 | 47.40 |
| Navi & Mani + *Multi Objects* + *Optimal Path with 2 Arms* | 72.04% | 74.21% | 28.12 | 28.51 | 5.60 (66.07%) | 5.61 (64.32%) | 38.25 | 39.17 |
| Navi & Mani + *Multi Objects* + *Temp Dependency* | 58.60% | 60.35% | 43.43 | 41.04 | 6.00 (64.05%) | 5.96 (62.71%) | 51.38 | 45.52 |

# Case Study: An Example of a Spatial constrained task


Can you get the box on the wall shelf?


1: Walk to the livingroom

2: Walk to the wall shelf (1)

3: Walk to the wall shelf (2)

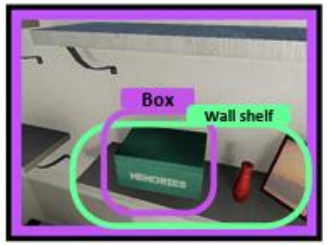
4: Walk to the wall shelf (3)


5: Walk to the bedroom

6: Walk to the wall shelf (4)

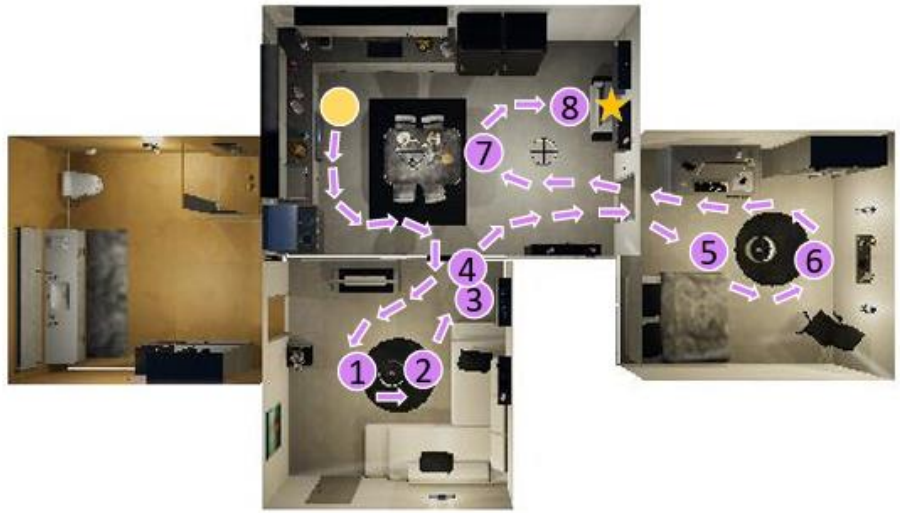

7: Walk to the kitchen


8: Walk to the box



Detected object


Final state




**Navi + Relation**

**Task: Can you get the box on the wall shelf?**

1: Walk to the livingroom
2: Walk to the wall shelf (1)
3: Walk to the wall shelf (2)
4: Walk to the wall shelf (3)
5: Walk to the bedroom
6: Walk to the wall shelf (4)
7: Walk to the kitchen
8: Walk to the box