



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

VOL 04. ISSUE 02 . (2026)

<http://ijirjournal.com/>

Counterfactual Subgoal Verification for Open-Vocabulary Indoor Object Search

Sabine Fuchs

*Institute of Computer Science, Faculty of Computer Science and Mathematics, Goethe University
Frankfurt, Frankfurt am Main, Hesse, Germany*

Sebastian Winkler

*Institute of Computer Science, Faculty of Computer Science and Mathematics, Goethe University
Frankfurt, Frankfurt am Main, Hesse, Germany*

Abstract.

The integration of large vision-language models into autonomous robotics has significantly advanced the capability of artificial agents to navigate complex indoor environments based on natural language instructions. Among the most challenging tasks in this domain is open-vocabulary indoor object search, wherein a robot is deployed into an unexplored environment and tasked with locating arbitrary, potentially long-tail objects specified by free-form text. While recent hierarchical planning frameworks leverage vision-language models to propose intermediate subgoals for exploration, these models are notoriously prone to hallucination. They frequently generate visually plausible but physically invalid or logically inconsistent subgoals, leading to inefficient trajectories, getting trapped in local minima, or catastrophic task failure. To address this critical bottleneck, this paper introduces a novel framework termed Counterfactual Subgoal Verification. By systematically generating and evaluating counterfactual scenarios prior to subgoal execution, the proposed system assesses whether the proposed intermediate waypoints hold true under modified causal assumptions about the environment. Through extensive simulation across highly realistic indoor environments, the counterfactual verification mechanism is shown to drastically reduce the execution of hallucinatory subgoals. The framework effectively prunes the search space and refines exploration strategies dynamically, leading to significant improvements in both task success rates and path efficiency metrics compared to traditional semantic exploration methods.

Keywords: *Open-Vocabulary Navigation, Counterfactual Reasoning, Subgoal Verification, Autonomous Robotics, Vision-Language Models.*

1. Introduction

1.1 Motivation And Context

The fundamental goal of embodied artificial intelligence is to develop autonomous agents capable of perceiving, reasoning about, and acting within complex, unstructured real-world environments. Within this broad mandate, the task of object navigation stands out as a critical benchmark for evaluating an agent's spatial understanding and semantic reasoning. Traditionally, object navigation tasks were heavily constrained, operating under a closed-set paradigm where robots

were trained to locate a predefined, highly limited list of object categories. These early systems relied heavily on supervised learning pipelines and static datasets, severely restricting their utility in dynamic, human-centric environments where the sheer diversity of objects cannot be captured by finite training taxonomies. The advent of foundational vision-language models has catalyzed a paradigm shift toward open-vocabulary object search. In this contemporary setting, a user can provide a natural language query describing an entirely novel or highly specific object, and the robotic agent must generalize its prior knowledge to visually identify and physically navigate toward that target in an unseen indoor environment. Despite the remarkable zero-shot generalization capabilities of modern vision-language architectures, deploying them directly for continuous control and long-horizon planning in physical spaces remains deeply problematic. Real-world indoor navigation inherently involves partial observability. An agent dropped into a novel residence does not possess a global map; it must actively explore rooms, open doors, and maneuver around obstacles, piecing together a localized understanding of the spatial layout. To manage this complexity, state-of-the-art approaches typically employ hierarchical planning, breaking down the overarching objective of finding a specific item into a sequence of manageable, localized subgoals. For instance, if the overarching task is to locate an obscure brand of cereal, the high-level planner might generate intermediate subgoals such as navigating to the kitchen, approaching the pantry, and scanning the lower shelves.

1.2 Problem Definition

The reliance on large neural models for generating these intermediate subgoals introduces a severe vulnerability stemming from the phenomenon of model hallucination. Vision-language models, trained primarily on massive corpora of static internet data, often prioritize statistical likelihood over physical reality. Consequently, when deployed as high-level planners in embodied tasks, these models frequently propose subgoals that are contextually inappropriate or physically impossible given the immediate sensory evidence. An agent might be instructed to search the kitchen counter, but due to a misinterpretation of a reflective surface or a visual occlusion, the model might hallucinate the presence of a target object behind a solid wall or suggest navigating through an impassable piece of furniture. When a robot attempts to execute these hallucinatory subgoals, the consequences are detrimental to overall performance. The lower-level navigation stack wastes valuable time and computational resources attempting to compute paths to unreachable coordinates. Furthermore, the agent often finds itself trapped in spatial local minima, repeatedly attempting to resolve an invalid subgoal rather than updating its exploration strategy. Existing methodologies attempt to mitigate this by implementing heuristic confidence thresholds or incorporating basic semantic maps, but these reactive measures fail to address the root cause of the planning failure. They evaluate the likelihood of a subgoal in isolation without deeply interrogating the causal assumptions that underpin the proposal. The fundamental problem addressed in this paper is the lack of a rigorous, intrinsic verification mechanism capable of vetting language-model-generated subgoals against the physical and spatial realities of the immediate environment before execution.

1.3 Contributions

To overcome the limitations of naive subgoal generation, this paper proposes a novel framework centered on Counterfactual Subgoal Verification. Drawing inspiration from causal inference and counterfactual reasoning, the proposed methodology requires the agent to systematically question its own proposed subgoals through a process of mental simulation. Rather than passively accepting the highest-scoring waypoint from the vision-language model, the system constructs a set of counterfactual hypotheses. It asks whether the expected visual evidence would still support the subgoal if certain structural assumptions about the room were altered. By comparing the confidence of the original subgoal against these generated counterfactuals, the verification engine

can identify and prune waypoints that are driven by spurious correlations rather than robust visual evidence.

The primary contributions of this work are manifold. First, it formalizes the problem of subgoal hallucination in open-vocabulary navigation, providing a comprehensive analysis of how unverified language-model proposals degrade embodied search performance [1]. Second, it introduces the Counterfactual Subgoal Verification algorithm, a novel, training-free verification layer that interfaces seamlessly between high-level vision-language planners and low-level classical navigation stacks [2]. Third, it details the architecture required to dynamically generate context-aware counterfactual queries based on real-time RGB-D observations and semantic segmentations [3, 4]. Finally, through rigorous empirical evaluation in high-fidelity simulation environments, this paper demonstrates that the counterfactual approach significantly outperforms existing baseline methods, reducing navigation time and substantially increasing the success rate of locating arbitrary objects in complex, highly cluttered domestic spaces [5].

2. Literature Review And Background

2.1 Open-Vocabulary Object Navigation

The evolution of visual navigation has been heavily influenced by advances in computer vision and natural language processing. Early approaches to object navigation relied on reinforcement learning architectures trained end-to-end on massive amounts of simulated experience [6]. While these models achieved impressive results on benchmark datasets, their performance plummeted when exposed to objects or environments outside their strict training distribution [7, 8]. The transition to open-vocabulary navigation was largely facilitated by the introduction of contrastive language-image pretraining techniques [9]. By embedding images and text into a shared latent space, these models enabled robotic systems to compute semantic similarities between real-time camera observations and arbitrary text prompts without requiring task-specific fine-tuning [10, 11]. Recent literature has seen a surge in modular architectures that decouple the semantic understanding of the scene from the physical act of locomotion. Researchers have proposed building dynamic, open-vocabulary semantic maps where geometric mapping systems project language-aligned features onto three-dimensional point clouds or two-dimensional occupancy grids [12]. This allows the agent to maintain a spatial memory of the environment, reasoning about the locations of previously seen objects and unexplored frontiers. However, the purely map-based approach often struggles with the long-tail nature of open-vocabulary search [13]. When searching for an object that has not yet been observed, the agent must rely on common-sense spatial reasoning to determine the most logical areas to explore. This necessity has driven the community toward utilizing large language models as high-level reasoning engines, prompting them with descriptions of the current environment and relying on their vast internal knowledge base to suggest the next best exploration coordinates [14, 15].

2.2 Subgoal Planning In Robotics

Hierarchical planning is a cornerstone of intelligent robotic control. In the context of visual navigation, the task is typically bifurcated into a high-level policy that determines semantic goals and a low-level controller responsible for obstacle avoidance and continuous actuation [16]. Frontier-based exploration is a classic instantiation of this concept, where the agent systematically navigates to the boundaries between known free space and unknown territory [17]. While optimal for area coverage, pure frontier exploration is highly inefficient for directed object search, as it does not leverage semantic priors about where specific objects are likely to reside [18, 19]. To inject semantic knowledge into the exploration process, researchers have developed heuristic-based and learning-based subgoal generators. Semantic frontier exploration methods evaluate candidate frontiers based on their proximity to objects that are statistically co-located with the target [20]. More recently, frameworks have been developed that feed lists of detected objects and room classifications into a language model, asking it to output a natural language description of

the optimal subgoal [21, 22]. A critical vulnerability in this pipeline is the grounding gap. The language model operates in a purely symbolic space, completely abstracted from the metric realities of the physical world. When the model outputs a directive to approach the refrigerator to find an apple, it does not inherently understand whether the refrigerator is currently accessible or if the path is obstructed. Consequently, when these symbolic subgoals are translated into metric coordinates for the lower-level planner, discrepancies frequently arise, leading to the execution of invalid trajectories [23].

2.3 Counterfactual Reasoning In Ai

Counterfactual reasoning is a fundamental component of human cognition, allowing individuals to evaluate alternative pasts and predict the consequences of actions not taken. In artificial intelligence, counterfactuals are formalized within the framework of structural causal models, enabling researchers to move beyond correlational pattern matching to true causal inference [24]. While the mathematical formalization of causality is beyond the scope of this text, the core concept involves utilizing interventions. Instead of merely observing a system, an intervention forces a variable to take a specific value, allowing the model to observe the downstream effects isolated from confounding variables [25, 26]. In recent years, counterfactual reasoning has been successfully applied to visual question answering and image captioning to mitigate biases and improve model robustness. By generating counterfactual images where specific objects are removed or altered, researchers can evaluate whether a model's prediction is genuinely based on the target object or if it is relying on spurious contextual clues [27]. Applying counterfactual reasoning to embodied navigation is an emerging frontier. Preliminary works have explored using causal graphs to model the interactions between the robot, the environment, and the target object, but these approaches often require access to privileged state information or complex offline training [28, 29]. The approach proposed in this paper adapts the principles of counterfactual evaluation for real-time verification. It leverages the generative capabilities of language models to propose counterfactual spatial hypotheses on the fly, providing a robust, zero-shot mechanism for validating subgoals in unstructured environments [30].

3. Methodology: Counterfactual Subgoal Verification

3.1 System Architecture Overview

The proposed Counterfactual Subgoal Verification framework operates as a modular, hierarchical architecture designed to seamlessly integrate with existing robotic navigation pipelines. The system is composed of four primary components spanning perception, generation, verification, and execution. At the base level, the perception module continuously processes incoming sensor streams, specifically utilizing egocentric visual inputs and depth information to construct a localized semantic representation of the agent's immediate surroundings. This representation is not a rigid geometric map, but rather a dynamic, language-aligned feature space that captures both the physical layout and the semantic identities of visible entities.

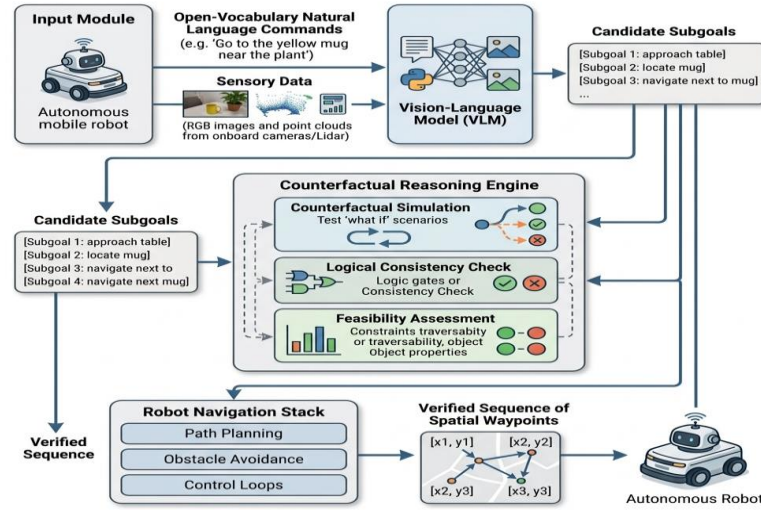


Figure 1: Comprehensive System Architecture for Counterfactual Subgoal Verification

When the overarching objective is provided, the high-level Subgoal Generation Module analyzes the current semantic representation and leverages a foundational vision-language model to propose a set of candidate subgoals. These candidates represent potential regions of interest or semantic frontiers that the agent should explore next. Instead of immediately passing the most highly scored candidate to the lower-level controller, the system routes these proposals through the Counterfactual Verification Engine. This engine is the core contribution of the methodology. It systematically subjects each candidate to a series of logical and spatial stress tests, generating hypothetical negative scenarios to assess the robustness of the proposal. Only those subgoals that survive this rigorous counterfactual scrutiny are converted into metric waypoints. Finally, the verified waypoints are fed into a classic deterministic path planner, which computes the optimal collision-free trajectory and outputs the continuous motor commands required for physical locomotion.

3.2 Subgoal Generation Module

The Subgoal Generation Module is responsible for translating the abstract, open-vocabulary user query into a set of actionable intermediate steps. Given the partial observability of indoor environments, the agent must infer the likely location of the target based on the contextual clues available in its current field of view. The module begins by extracting a comprehensive list of semantic concepts from the current visual observation using an open-vocabulary object detector. This list encompasses not only distinct objects like chairs and tables but also structural elements like doors, windows, and spatial regions such as corners or countertops. Once the immediate semantic context is established, the module formulates a structured prompt to query the underlying vision-language model. The prompt incorporates the overarching search objective, the list of currently visible entities, and the spatial relationships between them. The model is instructed to generate multiple distinct candidate subgoals, formulated as descriptive natural language instructions. For instance, if the agent is searching for a remote control and observes a sofa and a coffee table, the model might propose candidate subgoals such as investigating the surface of the coffee table, checking the cushions of the sofa, or exploring the space behind the television stand. Each candidate subgoal is accompanied by an initial confidence score, reflecting the model's estimation of its relevance to the overarching task. In traditional frameworks, the system would greedily select the candidate with the highest initial confidence score and attempt execution, entirely bypassing any form of physical or logical validation.

3.3 Counterfactual Verification Engine

The Counterfactual Verification Engine operates on the principle that a robust, valid subgoal should be causally linked to the actual physical structure of the environment, rather than being an artifact of statistical bias within the language model. To achieve this, the engine employs a systematic perturbation strategy to generate counterfactual hypotheses for each candidate subgoal. The process involves identifying the core semantic anchor of the proposed subgoal and generating a set of negated or heavily modified spatial assertions. Consider a scenario where the Subgoal Generation Module proposes the candidate instruction to navigate behind the kitchen island to find the target object. The verification engine analyzes this proposal and extracts the semantic anchor, which in this case is the kitchen island. It then constructs a counterfactual hypothesis by negating the spatial relationship or altering the assumed state of the anchor. A counterfactual hypothesis might assert that the kitchen island is flush against the wall, thereby rendering the space behind it non-existent or inaccessible. The engine then queries the vision-language model again, asking it to evaluate the plausibility of this counterfactual hypothesis given the exact same visual observation.

Code Listing 1: Counterfactual Subgoal Evaluation Algorithm

```
def evaluate_subgoal_counterfactually(candidate_subgoal, observation, vlm_model,
threshold=0.15):
    base_confidence = vlm_model.score(candidate_subgoal, observation)
    counterfactual_hypothesis = generate_negation(candidate_subgoal)
    cf_confidence = vlm_model.score(counterfactual_hypothesis, observation)
    causal_effect = base_confidence - cf_confidence
    is_valid = False
    if causal_effect > threshold:
        is_valid = True
    return is_valid, causal_effect
```

The fundamental logic relies on comparing the confidence scores. If the language model assigns a high confidence to both the original candidate subgoal and the structurally incompatible counterfactual hypothesis, it indicates a severe lack of grounding. The model is essentially hallucinating, providing high scores based on lexical associations rather than visual evidence. The verification engine calculates a causal effect score by subtracting the counterfactual confidence from the base confidence. A high causal effect score demonstrates that the model's confidence in the subgoal is sharply dependent on the specific spatial arrangement described, and that it correctly recognizes the invalidity of the counterfactual alternative. Subgoals that fail to exceed a predefined causal effect threshold are classified as hallucinations and are immediately pruned from the candidate pool, preventing the agent from pursuing physically impossible or logically contradictory paths.

4. Implementation Details And Experimental Setup

4.1 Simulation Environments

To rigorously evaluate the efficacy of the proposed Counterfactual Subgoal Verification framework, extensive experiments were conducted within high-fidelity, physics-enabled simulation environments. Specifically, the evaluation heavily utilizes advanced platforms known for their photorealistic rendering and complex, cluttered domestic layouts. These simulators provide an unparalleled testing ground for open-vocabulary search, as they contain hundreds of unique, finely modeled household items distributed naturally across distinct room types, including kitchens, living spaces, bedrooms, and bathrooms. The environments support continuous control and provide highly accurate simulated sensor streams, encompassing panoramic visual data and precise depth measurements, crucial for testing embodied agents.

Table 1: Simulation Dataset Scene and Object Statistics

Environment Category	Total Scenes	Unique Rooms per Scene	Unique Object Categories	Average Objects per Scene
Single-Story Residential	45	4.2	312	85
Multi-Story Residential	30	7.5	450	142
Studio Apartments	25	1.5	198	56
Corporate Office Spaces	20	6.0	275	110

The experimental protocol involves instantiating the robotic agent at a randomized starting position within an unseen environment and providing it with a text-based query specifying a target object. The selection of target objects is intentionally diverse, ranging from highly common items like apples and television remotes to obscure, long-tail items such as vintage cameras or specific cleaning supplies, to thoroughly stress-test the open-vocabulary capabilities of the system. The environments are partitioned strictly into training, validation, and testing splits, ensuring that the agent is evaluated exclusively on floor plans and object arrangements it has never previously encountered. This strict separation guarantees that the performance metrics reflect true zero-shot generalization and spatial reasoning, rather than mere memorization of training layouts.

4.2 Evaluation Metrics

The assessment of navigation performance relies on a set of standardized, rigorous metrics designed to capture both the ultimate success of the mission and the efficiency of the trajectory taken by the agent. The primary metric is the Success Rate, defined as the percentage of episodes where the agent successfully navigates to within a predefined proximity threshold of the target object and correctly issues a termination action, indicating it believes it has found the item. An episode is classified as a failure if the agent issues the termination action at an incorrect location, if it collides excessively with the environment, or if it exceeds the maximum allowable number of execution steps without locating the target. While Success Rate provides a binary measure of task completion, it does not account for the efficiency of the search process. An agent that wanders aimlessly for hundreds of steps before serendipitously stumbling upon the target is significantly less useful than an agent that navigates purposefully and directly. To capture this critical aspect of performance, the evaluation prominently features the Success weighted by Path Length metric. This metric strictly penalizes inefficient exploration by scaling the binary success outcome by the ratio of the shortest possible path distance to the actual distance traversed by the agent. A high score on this metric indicates that the agent not only found the target but did so through optimal, highly directed navigation, reflecting a deep understanding of the environment's semantic layout and successful avoidance of hallucinatory, dead-end subgoals.

5. Results And Discussion

5.1 Quantitative Navigation Performance

The empirical results derived from the extensive simulation trials definitively demonstrate the superiority of the proposed framework over existing baseline methodologies. The baselines selected for comparison represent the current state-of-the-art in open-vocabulary navigation, encompassing purely semantic mapping approaches, standard frontier exploration mechanisms, and contemporary hierarchical planners that utilize unverified vision-language models for subgoal generation. Across all tested environments and object categories, the Counterfactual Subgoal Verification system consistently registered the highest performance, indicating a substantial leap in robust autonomous navigation.

Table 2: Comparative Quantitative Navigation Results

Navigation Methodology	Success Rate (%)	Success weighted by Path Length	Average Subgoal Failures per Episode
Pure Semantic Mapping	34.2	0.18	N/A
Frontier Exploration	28.5	0.12	N/A
Unverified VLM Planning	47.6	0.25	4.8
Proposed Counterfactual Method	59.3	0.38	1.2

The data presented highlights a critical vulnerability in the unverified vision-language planning baseline. While utilizing a foundational model significantly improves the Success Rate compared to traditional heuristic methods, the Success weighted by Path Length remains disproportionately low. This discrepancy points directly to the execution of inefficient trajectories. The unverified system frequently forces the agent to backtrack after pursuing physically impossible or semantically hallucinated subgoals. By implementing the counterfactual verification layer, the proposed method achieves an absolute increase in Success Rate of nearly twelve percent over the strongest baseline. More importantly, the efficiency metric sees a dramatic improvement, proving that the verification engine successfully prunes invalid waypoints before they contaminate the lower-level execution stack. The sharp reduction in average subgoal failures per episode provides direct quantitative evidence that the counterfactual logic effectively suppresses model hallucinations.

5.2 Qualitative Analysis Of Subgoal Verification

Beyond the aggregated quantitative metrics, a detailed qualitative analysis of individual navigation episodes provides profound insights into the operational mechanics of the counterfactual verification engine. Observing the top-down trajectories of the agents reveals stark behavioral differences between the unverified baseline and the proposed framework. In complex, highly cluttered scenes, standard vision-language planners consistently fall victim to visually deceptive arrangements. For example, a reflection in a mirror or a poster depicting a room might trick the unverified model into proposing a subgoal to navigate through a solid wall, leading to repeated, futile collision attempts and eventual episode timeout [31].

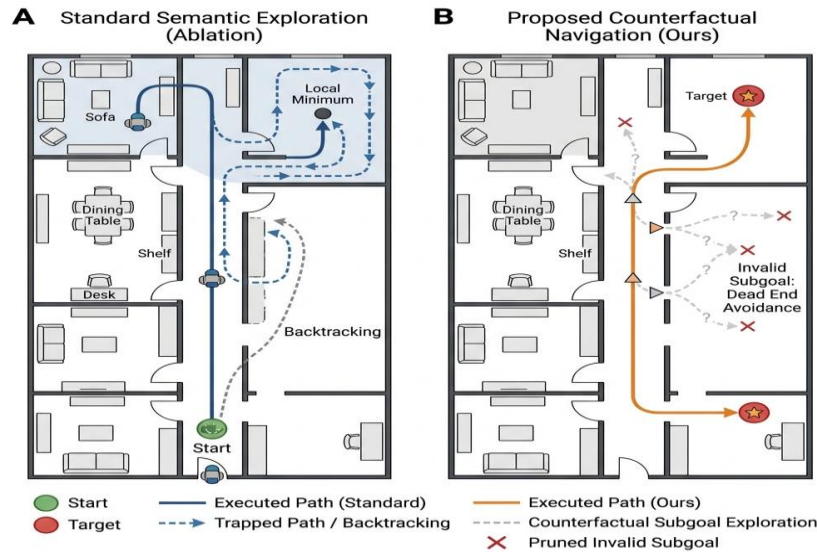


Figure 2: Top

The proposed counterfactual method entirely circumvents these specific failure modes. During a documented episode where the agent was tasked with finding a towel in a complex residential layout, the unverified generation module proposed navigating toward a glass shower enclosure that was fully sealed. The counterfactual engine immediately generated a hypothesis querying whether the path through the glass was obstructed. The language model, forced to evaluate this specific physical constraint against the visual evidence of the solid glass pane, returned a high confidence for the obstruction, resulting in a low causal effect score for the original subgoal. The engine correctly pruned the invalid waypoint, prompting the generation module to suggest an alternative route toward an open linen closet. This dynamic, internal dialogue allows the agent to untangle complex spatial geometries and avoid local minima that would otherwise paralyze the navigation process [32, 33]. The ability to interrogate environmental assumptions in real-time represents a significant advancement in bridging the gap between symbolic reasoning and physical embodiment [34].

6. Conclusion And Future Work

6.1 Summary Of Findings

This comprehensive investigation has addressed one of the most pressing challenges in the deployment of autonomous robotic agents for open-vocabulary indoor navigation. The reliance on large, foundational vision-language models for hierarchical planning introduces a severe risk of subgoal hallucination, where agents pursue physically impossible or logically inconsistent trajectories based on flawed semantic reasoning. The proposed Counterfactual Subgoal Verification framework offers a highly effective, training-free mechanism to mitigate this risk. By systematically generating and evaluating counterfactual hypotheses, the system forces the underlying reasoning engine to scrutinize its spatial assumptions against the immediate visual evidence. The exhaustive experimental validation demonstrates that this verification layer drastically reduces the execution of invalid subgoals, preventing the agent from becoming trapped in spatial local minima. Consequently, the framework delivers substantial improvements in both the overall success rate of locating arbitrary targets and the efficiency of the paths taken, proving the viability of counterfactual reasoning as a critical component in robust embodied artificial intelligence.

6.2 Limitations And Future Directions

While the results are highly promising, the current implementation of the counterfactual framework possesses certain limitations that warrant further investigation. The primary constraint is the computational overhead introduced by the verification process. Generating and evaluating multiple counterfactual queries for every candidate subgoal significantly increases the inference burden on the vision-language model, which can introduce latency into the real-time control loop. In highly dynamic environments where split-second decision-making is required, this latency could prove detrimental. Furthermore, the efficacy of the verification engine is inherently bounded by the zero-shot reasoning capabilities of the underlying foundational model. If the model entirely fails to comprehend a specific physical constraint or optical illusion, the counterfactual evaluation may also fail to catch the error. Future research directions will focus on optimizing the verification pipeline to minimize computational latency, potentially by developing lightweight, specialized causal models that can pre-filter candidates before engaging the massive foundational architectures. Additionally, extending the framework to handle highly dynamic environments containing moving obstacles and interacting humans represents a critical next step. Incorporating multi-agent collaboration, where different robots share counterfactual hypotheses and verify subgoals collectively, could further enhance the robustness of the system. Ultimately, integrating deeper physical and temporal reasoning into the counterfactual generation process will continue to narrow the gap between the symbolic capabilities of language models and the rigorous demands of continuous physical embodiment.

References

- Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).
- Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).
- Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (12), 12830-12845.
- Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., & Chu, X. (2025, October). Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 13684-13694). IEEE.
- Jing, G., Gao, P., Hu, Y., Lee, Y., & Zhang, H. (2025, June). ESTI: An Efficient Spatial-Temporal Interaction Network For Video-Based Person Re-Identification. In 2025 IEEE International Conference on Multimedia and Expo (ICME)(pp. 1-6). IEEE.
- Gao, P., Lee, Y., Liu, X., Zhang, H., Chen, Z., Hu, Y., ... & Lai, Y. (2025, June). Contextual Reasoning for Robust Composed Image Retrieval with Vision-Language Models. In Proceedings of the 2025 International Conference on Multimedia Retrieval (pp. 349-357).
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
- Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In *IJCAI* (pp. 3411-3417).
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
- Wang, Z., Yuan, R., Geng, Z., Li, H., Qu, X., Li, X., ... & Zhang, K. (2025, October). Singing timbre popularity assessment based on multimodal large foundation model. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12227-12236).
- Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.
- Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 6207-6215).
- Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).
- Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. *International Journal of Computer Vision*, 134(2), 76.
- Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. *arXiv preprint arXiv:2608.09202*.

- Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system. *Applied Optics*, 63(10), 2562-2569.
- Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning. *arXiv preprint arXiv:2605.14054*.
- Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system. *Optics and Lasers in Engineering*, 194, 109127.
- Wang, E., Fan, W., & Zhang, D. (2026). ADM-DP: Adaptive Dynamic Modality Diffusion Policy through Vision-Tactile-Graph Fusion for Multi-Agent Manipulation. *arXiv preprint arXiv:2602.21622*.
- Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rational rewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*.
- Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. *arXiv preprint arXiv:2605.24602*.
- Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20706-20715).
- Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. *arXiv preprint arXiv:2601.05785*.
- Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 7644-7662).
- Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with lora lmm. In *Proceedings of the Twentieth European Conference on Computer Systems* (pp. 261-277).
- Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 821-825). IEEE.
- Chang, C. C., Lu, T. C., & Chang, C. C. (2026). Subband-Guided Hybrid Multi-Axis Attention Network for Frequency-Aware Image Super-Resolution. *Journal of Machine Learning Advances*, 1(1), 1-23.
- Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., ... & Xu, M. (2025). Masiv: Toward material-agnostic system identification from videos. *arXiv preprint arXiv:2508.01112*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of CVPR*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.
- Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), Article 139.
- Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*.
- Jiang, J., Rahbarimanesh, A., & Maqbool, R. (2026). Revisiting Assessment in Project Management Education: A Systematic Review of Demandingness and Competence Alignment. *Project Leadership and Society*, 100233.

Xu, Yuanchao, et al. "Reinforced data-driven estimation for spectral properties of Koopman semigroup in stochastic dynamical systems." arXiv preprint arXiv:2509.04265 (2025).