IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

GAIA: Generating Task Instruction Aware Simulation Grounded in Real Contexts using Vision-Language Models

Dogyu Ko†, Chanyoung Yeo†, Daeho Kim, *Student Member, IEEE*, Jaeho Kim, and Hyoseok Hwang, *Associate Member, IEEE*

***Abstract*—Enabling robots to interact effectively with the real world requires extensive learning from physical interaction data, making simulation crucial for generating such data safely and cost-effectively. Despite the advantages of simulation, manual environment creation remains a laborious process, motivating the development of automated generation approaches. However, the limitations of current automatic virtual scene generation approaches in bridging the sim-to-real gap and achieving task readiness necessitate the creation of automatically generated, realistic, and task-ready virtual scenes. In this paper, we propose GAIA, a novel methodology to automatically generate interactive, task-ready simulation environments grounded in real contexts from only a single RGB image and a task instruction. GAIA utilizes a pre-trained Vision-Language Model (VLM) without requiring explicit training, and jointly understands the visual context and the user's instruction. Based on this understanding, it infers and places necessary task-aware objects, including unseen ones to construct an interactive virtual environment that maintains real-scene fidelity while reflecting task requirements without additional manual setup. We show qualitative experiments that GAIA generates spaces consistent with user instructions, and quantitative results that policies learned within these GAIA-generated environments successfully transfer to target environments. Source code and supplementary materials are available at our project page https://sites.google.com/view/gaia-project-page.**

***Index Terms*—Simulation and Animation, Task and Motion Planning, Deep Learning for Visual Perception**

I. Introduction

MODERN Artificial Intelligence (AI) addresses complex challenges and sees broad adoption across various domains, from industrial manipulation [1], [2] to social applications [3], [4], supported by rich datasets [5], [6], computational power, and advanced machine learning algorithms [7].

Manuscript received: June, 30, 2025; Revised August, 27, 2025; Accepted September, 23, 2025.

This paper was recommended for publication by Abhinav Valada upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under Grant RS-2025-00564137, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) under Grant RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), and in part by Convergence security core talent training business support program under Grant IITP-2023-RS-2023-00266615). *(Corresponding author: Hyoseok Hwang.)*

†These authors contributed equally to this work.

The authors are with the Department of Software Convergence, Kyung Hee University, Yongin-si, Gyeonggi-do, 17104, Republic of Korea (e-mail: kodogyu@khu.ac.kr; ducksdud08@khu.ac.kr; kdh2769@khu.ac.kr; leokim51@khu.ac.kr; hyoseok@khu.ac.kr).

Digital Object Identifier (DOI): see top of this page.

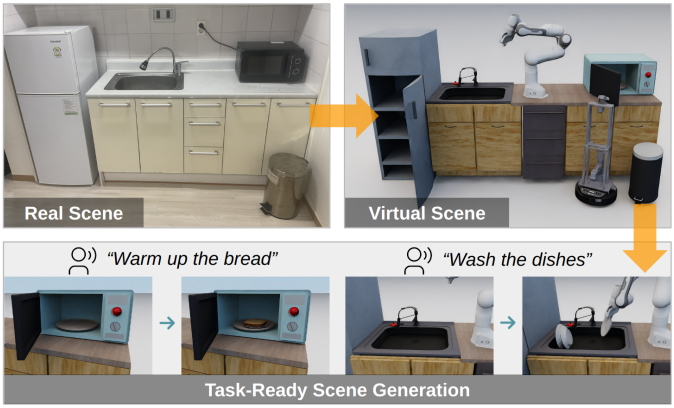


Fig. 1. GAIA generates task-ready simulation scene from a single RGB image and an abstract user instruction. It utilizes a VLM to generate virtual scene from a RGB image then based on the semantic information, it places the task object for agent policy learning.

To ensure general performance in real-world scenarios, pre-training on broad data has become instrumental for developing high-capability models [8]. Following this initial phase, task-specific fine-tuning adapts these general models, enabling them to achieve high performance in specialized downstream tasks [9]–[11]. Recently, Vision-Language Models (VLMs) [12], [13], trained on scalable online image and text data, effectively address complex high-level visual-language tasks such as visual reasoning effectively without task-specific finetuning.

Advancements in multimodal AI capabilities are stimulating research into embodied AI, where agents operate beyond digital boundaries to interact directly with the physical world [14]. The primary goal of embodied AI is to develop agents capable of understanding and acting within complex physical environments to accomplish sophisticated tasks, such as long-horizon navigation [15] and object manipulation [16] following natural language instructions. Achieving this requires extensive learning from physical interaction data. However, collecting real-world data poses significant challenges, including hardware expenses, safety risks, and the substantial time costs of collecting data [17].

As a cornerstone of embodied AI research, simulation enables robot agent interaction within virtual environments, addressing real-world data acquisition limitations. Its primary contribution lies in facilitating the generation of extensive, cost-effective, and realistic data for training agents on diverse tasks [15], [18]. Additionally, it offers rigorous testing under diverse conditions and ensures high reproducibility needed

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

for reliable verification. Nevertheless, creating high-fidelity simulation environments remains challenging. Manual creation of diverse, task-specific virtual scenes is indeed a time-consuming and expertise-intensive bottleneck for rapid robot adaptation [19], [20]. Consequently, methods for automatically generating varied, task-relevant simulation environments are crucial to overcome these limitations and enhance agent learning and adaptability.

Research on automatic simulation environment generation largely follows two approaches: text-guided and image-based methods. Text-guided approaches [19], [21], [22] offer flexibility, generating diverse environments from task instructions, yet are unable to accurately capture the real-world deployment context, resulting in a sim-to-real gap. In contrast, image-based approaches [23], [24], provide high visual fidelity by creating virtual scenes, potentially improving sim-to-real transfer. However, these approaches often lack task-readiness because they only replicate visible objects, requiring users to manually configure scenes for robot learning. Therefore, a primary goal is to merge task instruction with image grounding for automatic, real-scene based, task-ready virtual scene creation. Considerable room for research remains in enabling the automatic and meaningful placement of task-aware objects within the realistic virtual scene to effectively support robot learning.

To address the challenge of automatically placing task-aware objects meaningfully while considering a real-world context, we present **GAIA**, a novel methodology leveraging VLM for automatic, task-ready virtual scene generation (see Fig. 1). GAIA utilizes the pre-trained VLM requiring no additional learning to jointly reason about visual-language information from a single RGB image and an abstract task instruction. This inherent zero-shot capability allows GAIA to handle diverse scenes and instructions, configuring virtual environments appropriately for robot learning tasks. Our proposed framework operates in three stages: 1) Instruction Understanding (Sec. III-A), comprehending the user's request relative to the scene context to generate task-ready virtual scene scenarios; 2) Object Configuration (Sec. III-B), identifying, retrieving, and arranging necessary objects meaningfully; and 3) Task-Aware Scene Generation (Sec. III-C), resulting in diverse, task-ready virtual scenes based on the initial real scene context, suitable for robot policy learning. In summary, our main contributions are as follows:

- We propose a novel methodology, GAIA, that automatically generates various task-ready virtual scenes by leveraging a VLM to understand task instructions from a single RGB image.
- Leveraging a pre-trained VLM enables task-aware objects to be placed in a contextually appropriate way that is suitable for robot policy learning across a wide range of general scene environments.
- Experimental results demonstrate that GAIA effectively generates task-ready scenes that facilitate the learning of valid, task-achieving robot policies.

II. Related Work

A. Virtual Scene Generation for Robotics

Research in virtual scene generation encompasses approaches focusing on scene synthesis [25], [26], often resulting in static environments that limit physical robot interaction. To enable interaction, recent studies [27], [28] incorporate articulated objects and integrate with interactive 3D simulation platform. Architect [29] presents another approach, utilizing diffusion based inpainting for complex layouts, though it often requires manual steps for policy learning integration. Another line of research involves text-to-scene generation [21], [22]. Leveraging foundation models, these approaches offer convenient creation of diverse environments suitable for policy learning but often struggle with the sim-to-real gap. Methods focusing on realism reconstruct environments similar to the real world [30], [31], but making these reconstructed scenes interactable usually requires user intervention. Alternatively, ACDC [24] retrieves geometrically similar 3D assets based on objects in real images, demonstrating effective sim-to-real transfer. However, it struggles to generate objects not present in the input image and requires user guidance to tailor environments for specific robot tasks. To overcome these limitations, our proposed method uniquely combines the strengths of text understanding and real image grounding.

B. Large Vision-Language Models for Robotics

Recent remarkable advancements in Large Language Models (LLMs) [32], [33] have revolutionized natural communication between humans and computers. Building on this success, VLMs [12], [13], extend LLM capabilities with visual understanding. These models show outstanding performance and versatility in complex multimodal tasks like Visual Question Answering (VQA) and image captioning. In robotics, advanced visual-language understanding of VLMs is being leveraged for direct robot action control [34], [35]. In addition, VLMs have shown potential applications in various high-level robotics domains such as reward function design [36], task planning [37], error correction [38], optimization strategies [39] and retrieving appropriate action data [40]. One particularly relevant VLM application for our research is the effective generation of virtual environments for robot policy learning. Previous works [19], [41] have proposed creating virtual environments suitable for robots performing daily tasks. Specifically, RoboGen [19] is the relevant work to our study in that it automatically generates simulation environments for robot learning utilizing large models. However, this approach focuses on ensuring data diversity by generating virtual environments from scratch using only text prompts, which may have limitations in terms of the sim-to-real gap that occurs when transferring policies to specific real environments. Another approach, Eurekaverse [42], focuses on the step-by-step design of virtual spaces tailored for curriculum-based locomotion policy learning. Our research comprehensively leverages the language understanding, spatial reasoning, and context-based retrieval capabilities inherent in VLMs. The goal is to generate virtual spaces where robots can effectively learn

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

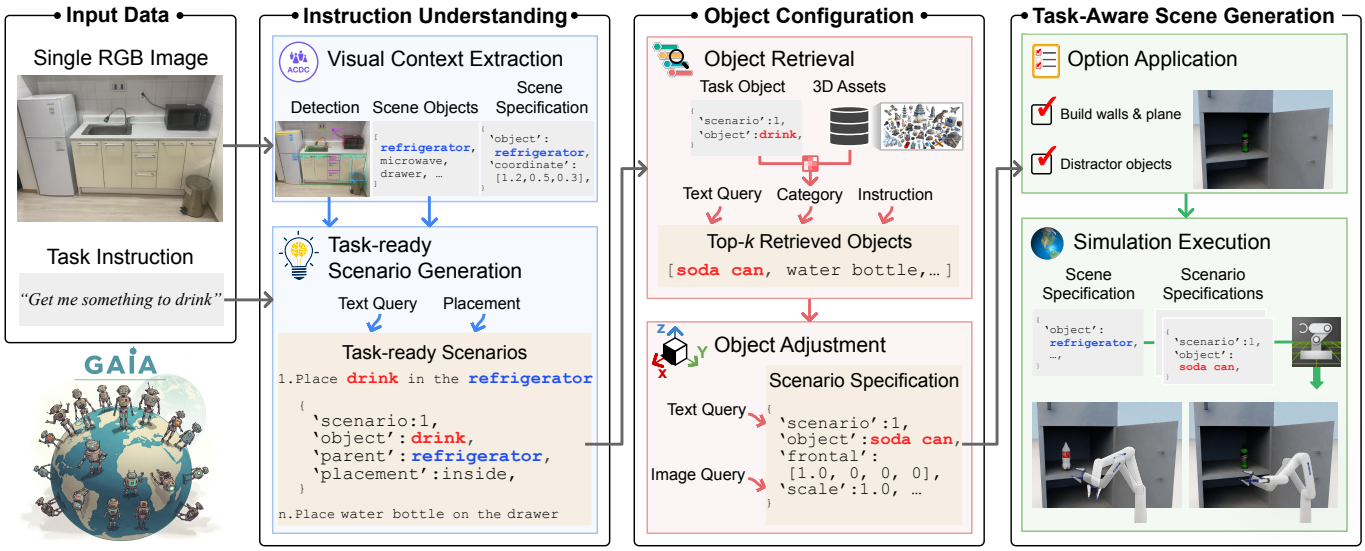


Fig. 2. Overview of the GAIA framework. GAIA uses a VLM to interpret spatial context from an image and semantic intent from a task instruction. Based on this understanding, it automatically retrieves the necessary 3D assets to build an interactive simulation ready for embodied AI.

complex daily tasks, thereby facilitating direct policy learning and seamless transfer to real-world applications.

III. Methodology

The primary goal of our research is to enable the creation of interactive, scene-level virtual environments suitable for embodied AI. Our proposed method GAIA focuses on task-aware scene generation, leveraging a pre-trained VLM [12]. This approach utilizes both a single real-world RGB image and the user's task instruction. Fig. 2 presents our framework, proceeding through three stages: First, **Instruction Understanding** processes the inputs to understand the user's task instruction and the necessary characteristics for task-ready scenarios. Second, **Object Configuration** retrieves, adjusts, and places the necessary task-aware objects in the virtual scene. Finally, **Task-Aware Scene Generation** runs optional steps and creates the complete interactable virtual scene.

A. Instruction Understanding

The objective of this stage is to interpret the input task instruction considering the visual context provided by the single RGB image. This defines the specifications for a task-ready scenario that is suitable for subsequent agent learning and execution. This process utilizes the VLM to analyze semantic properties and object relationships depicted in the image, jointly reasoning about this visual information in conjunction with the textual task instruction. For example, given the instruction "*Get me something to drink*" and an input image containing a refrigerator, the VLM leverages its inherent world knowledge that refrigerators typically contain beverages. Based on this reasoning, the generation process ensures that the resulting task-ready scenario includes appropriate drinkable items, such as water bottles or juice cans, placed inside the refrigerator.

Visual Context Extraction. Inspired by ACDC [24], this step processes the input RGB image to extract essential visual context required for subsequent task-aware scenario generation. We utilize GroundedSAM-v2 [11] to detect and segment visible objects, including walls and the floor, producing an annotated image and a list of scene objects. These are then used to extract relevant semantic information associated with each detected object. Subsequently, for each detected object, the best matching 3D asset is retrieved from a library [43] based on DINOv2 [10] embeddings. Pixel-wise depth is also estimated from the image via Depth Anything-v2 [9] to adjust the scale of retrieved assets for improved realism. Finally, all extracted information is compiled into a structured visible scene specification for simulation generation.

Task-Ready Scenario Generation. This step employs the annotated image, scene objects list, text query, and task instruction as inputs to generate task-ready scenarios. The VLM identifies the necessary task-aware objects by interpreting the task instruction within the given visual context. This may involve inferring objects not visible in the original image based on world knowledge. For instance, processing "*Get me something to drink*" with a visible refrigerator leads the VLM to specify the addition of beverage items like water bottles or juice cans. Jointly, identifying the required objects, the system determines their placement. This involves selecting a suitable spatial relationship {*Left, Right, Above, Below, Front, Back, Inside*} relative to a parent object (e.g. visible refrigerator or other task-aware object). The VLM selects this parent object from existing scene elements to act as a spatial landmark, defining the target location based on the chosen relationship. The final output is a structured specification listing of the identified task-aware objects and their determined target spatial configurations.

B. Object Configuration

Building upon the previously generated task-ready scenario, this stage retrieves specific objects suitable for the task from a predefined 3D asset library [43]. Subsequently, the intrinsic properties of the retrieved objects, specifically scale and orientation, are adjusted in accordance with the real scene scale.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

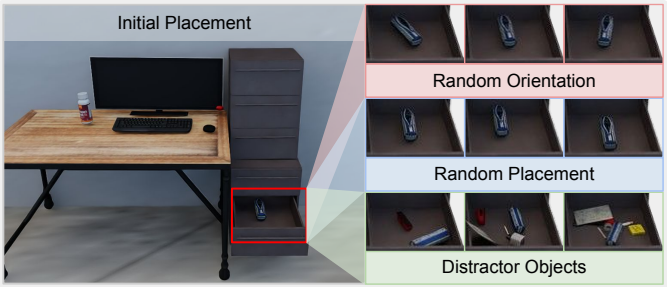


Fig. 3. Object placement augmentation. Left: initial placement of target object, Right-top: random orientation, Right-middle: random placement, Right-bottom: distractor objects.

We utilize the CLIP [44] encoder to guide appropriate 3D asset selection from the predefined 3D asset library, along with the VLM for semantic reasoning, contextual understanding, and scale adjustment. This process completes the specification of a structured scenario for task-aware objects that is both semantically and physically plausible and ready for agent task execution.

Object Retrieval. Given the scenario specification for a required task-aware object type, this step retrieves the top-k matching 3D assets from a predefined library [43]. Initially, GAIA selects N candidate categories from the library exhibiting the highest semantic similarity to the required object type, measured using cosine similarity between their respective CLIP features. Subsequently, for all objects within these selected categories, GAIA concatenates their rendered images with corresponding indexing information. Then, the VLM processes a prompt containing the full task and environmental context, evaluating the concatenated information to score each candidate asset based on its suitability. The output of this step comprises the top-k ranked 3D assets identified as most appropriate.

Object Adjustment. The selected top-k objects often have distinct coordinate systems, which complicates direct placement and interaction in the virtual scene when using only default orientations or manual rules. To address this, we first adjust the object's orientation. For each selected object, we generate four candidate views by rotating it incrementally by 90°around the z-axis of its object coordinate system, capture images of these orientations, index them, and concatenate the images. We then provide another query, including relevant input data such as the annotated image and scene object list, to select the single most suitable orientation from the four candidates for the intended task execution. This selection ensures that the object is oriented in a way that is functional for the task, rather than merely presenting its default front view. This contributes to a scene configuration that is aware of the task.

Additionally, to augment the object's orientation, GAIA provides an option for random orientation that samples an angle around the z-axis within a bounded range centered around the semantically selected orientation, which is illustrated in Fig. 3 right-top. The range is user-configurable, allowing users to increase the diversity of object placements while preserving semantic plausibility. Finally, to ensure the object has an appropriate size, we first determine the bounding box dimensions of the object used in the simulation. Based on this, we query the VLM for the typical size of that object, and then calculate a scaling factor accordingly.

C. Task-Aware Scene Generation

Object Placement. This process determines the optimal placement for the task-aware objects. When the *Above* relationship is selected, we first establish a 3×3 grid on the parent object's top surface from an RGB image. We then use a VLM to estimate a probability of semantic suitability for placing the object at each grid cell's center point. Based on these scores, an initial Gaussian heatmap of placement likelihood is generated. To avoid collisions, this map is then refined by masking out areas occupied by pre-existing objects, considering the bounding boxes of both the obstacles and the new object. The location with the highest probability in the final, refined map is selected as the target position.

To prevent excessive spatial variation, we allow users to set a maximum randomization range. For all other spatial relationships except the *Above* and *Inside* relationships, the task-aware object is centered on the corresponding face of the parent object's bounding box, adjusted for its own dimensions. This overall method integrates semantic reasoning with geometric constraints to facilitate intelligent object placement.

Option Application. In addition to the main structure, GAIA provides two options to generate more realistic scenes and thereby reduce the sim-to-real gap. First, the Build Background option constructs wall and floor planes based on the input image. Second, the Distractor Objects option places objects that are semantically appropriate for the task instruction, which not only enhances scene realism but also improves the robustness of the learned policy.

When building background, the system utilizes segmentation masks and depth images, all obtained during the Visual Context Extraction stage. These are used to generate point clouds, from which the normal vectors of the corresponding planes are estimated using Principal Component Analysis (PCA). To further enhance the realism of the scene, textures are extracted for each surface in the input image by computing a homography to obtain a fronto-parallel view, followed by rectifying both the RGB image and segmentation mask. This process enables realistic texture extraction.

For distractor object placement, if the target object is assigned an *Inside* spatial relationship, GAIA populates the parent object with semantically appropriate distractors (see Fig. 3 right-bottom). A VLM identifies these distractors based on the scene context. They are then retrieved from a 3D asset library and configured. The objects are placed via physics-based dropping to ensure settling, followed by the target object in the same manner. This results in realistic, cluttered configurations that improve policy robustness in complex environments.

Simulation Execution. In the final Task-Aware Scene Generation stage, the initial scene specification from Visual Context Extraction stage is combined with the task-aware objects specification generated during Object Placement stage. This comprehensive specification is then loaded into physics simulator NVIDIA Omniverse Isaac Sim [45], creating the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

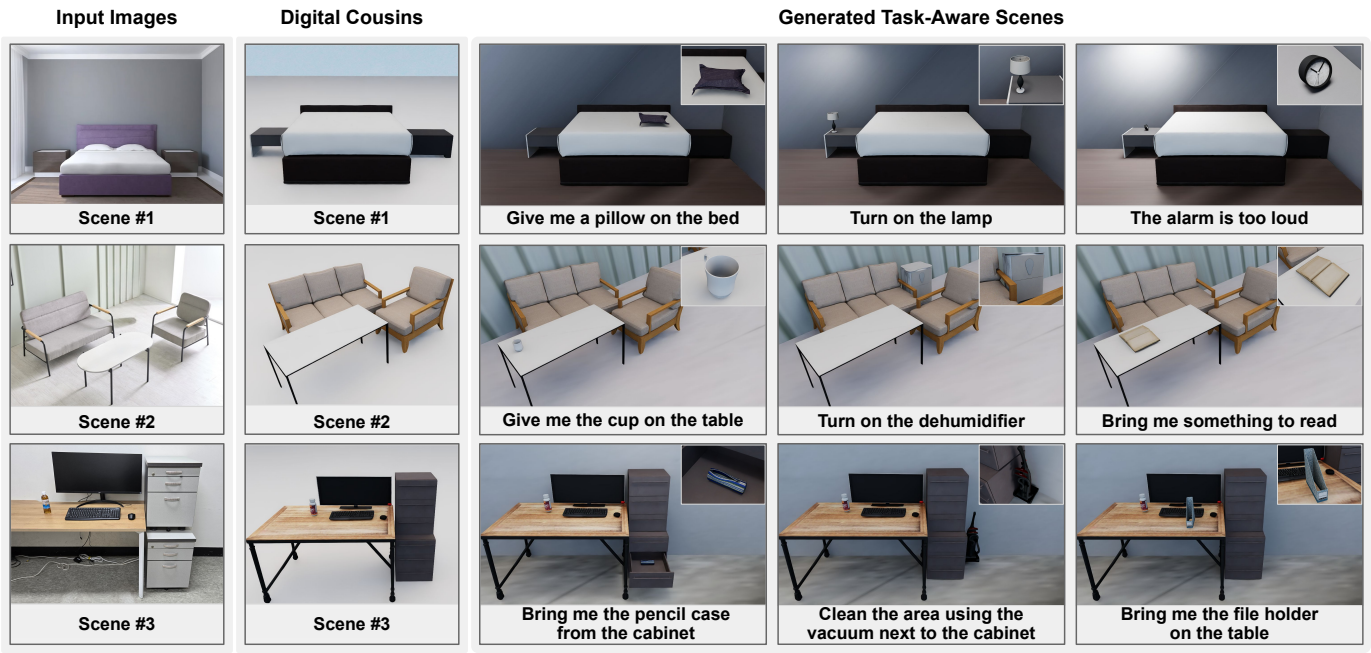


Fig. 4. Examples of task-aware scenes generated by GAIA from real RGB images. Given an input scene, a digital cousin is created in a virtual environment, and then task-aware scenes are constructed for various task instructions. Top-right insets offer zoomed-in views of the generated task-aware objects.

interactive virtual environment. Notably, virtual scenes from the GAIA framework are immediately usable by embodied agents for interaction and collecting necessary data for the instructed task, requiring no extra manual setup. Furthermore, following the digital cousin [24] concept, data gathered in simulation supports direct real-world transfer.

IV. Experiments

Our experiments address the following three questions:

1. How effectively does GAIA generate task-ready scenes for robotic tasks, given only a single RGB image and a task instruction?
2. Are the virtual scenes generated by GAIA inherently suitable for training effective robot agent policies without requiring further manual configuration?
3. How effective is the sim-to-real transfer of robot policies trained within GAIA-generated virtual scenes to real-world scenarios?

A. Evaluation Metrics

To evaluate the fidelity of our generated virtual scenes to the original input, their suitability for the instructed task, and the resulting policy performance, we adopt the following evaluation protocol [29]. First, CLIPScore [46] measures the similarity between CLIP embeddings of rendered views from the generated scene and the input textual instruction. Second, BLIPScore computes alignment between rendered scene views and the instruction using BLIPv2's [47] image-text matching head. Third, VQAScore [48] uses the probability of 'Yes' answer from a VQA model queried with the rendered view and instruction about accurate setup reflection: "*Does the scene accurately reflect the setup, requested in the instruction?*" Finally, success rate measures the percentage of successful task completions.

TABLE I
Text-to-Image Evaluation for Generated Virtual Scenes

Method	CLIP ↑	BLIP ↑	VQA ↑
RoboGen [19]	0.61273	0.23545	0.65695
Ours	<u>0.61661</u>	<u>0.31866</u>	**0.87579**
Heuristic (Oracle)	**0.64120**	**0.34709**	<u>0.85999</u>

B. Results on Virtual Scene Generation

Qualitative Results of Virtual-Scene Generation. Addressing our Question 1, we qualitatively evaluated the task-ready virtual scenes generated by GAIA from input images and various task instructions. Fig. 4 illustrates representative results using web-sourced images and our own environment images as input, which are displayed in the first column. The second column shows the results from the ACDC [24] method used to create a digital cousin of the environment, and the subsequent columns present the task-ready virtual environments generated by GAIA for the same virtual scene with different task instructions. For these qualitative results, we enabled only the Build Background option to clearly assess GAIA's core capabilities and minimize visual obstruction from other objects.

As the results demonstrate, GAIA successfully interprets the intent behind task instructions to introduce semantically relevant objects. Its sophisticated spatial reasoning is across the overall results. For instance, objects like the cup and pillow in the first row are placed in *Above* spatial relationships by utilizing a VLM. In the third row office scene, GAIA also successfully put a file holder on a cluttered desk, showing its collision avoidance ability. Moreover, even when a parent object wasn't explicitly defined in the instructions, the dehumidifier in the second row was placed between sofas, and the alarm in the first row was placed on a nightstand,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Fig. 5. Results of Task-Aware Scene Generation for sim-to-sim policy learning. Top: RoboGen [19], Bottom: Ours. For each task instruction, we find that our method generated a more semantically appropriate scene compared to the baseline. The input RGB image for our model is illustrated top-right.

both accurately. As seen with the pencil case in the third row, the model capably handles other spatial relationships, like *Inside*. These results demonstrate that our proposed method effectively considers the scene context and task instruction to generate plausible task-ready virtual scene without any manual intervention.

Quantitative Results of Virtual-Scene Generation. To quantitatively address our Question 1, We compared our generated virtual scenes against two baselines: RoboGen [19], and a Heuristic baseline. For the Heuristic baseline, a human operator was given a target image and task instruction and then constructed a similar, executable scene. To ensure a fair comparison, the operator was restricted to the same 3D asset dataset [43] used by GAIA. For consistency with the qualitative analysis, GAIA was evaluated under the same settings. For the evaluation, we curated 20 household tasks from ManiSkill-HAB [49], building upon the benchmark's Pick and Place skills and augmenting them with a Move skill to cover a broader range of instructions. The evaluation prompts were generated using the DSG [50] approach, first as questions and then reformulated into declarative sentences for text-image metrics.

As shown in Table I, GAIA achieves a comparable CLIP-Score to the heuristic baseline. Similarly, the BLIPScore and VQAScore of GAIA are also close to or even surpass those of the heuristic baseline, indicating that our method performs competitively compared to carefully crafted human-designed scenes. This demonstrates its ability to simultaneously satisfy the demands of task relevance. In contrast, RoboGen tends to exhibit lower scores, primarily due to its limited ability to generate task-relevant objects. For instance, it generated an incorrect object in place of an apple in "*Place the apple from the fridge into the bowl*" or applied incorrect scaling to a bottle, resulting in an appearance that did not resemble a bottle in the scene. These results quantitatively validate GAIA's superior capability in producing virtual environments that are highly faithful to text inputs, confirming their suitability as effective task-ready scenes.

Ablation Study. We proposed a method to place task-aware objects in a manner that is semantically plausible within a given RGB scene. For our experiment, we evaluated the performance of our approach with and without Front View estimation, Object Resizing, and Build Background using the same evaluation protocol as in the quantitative results, based on text-image metrics for the 9 task instructions from Fig. 4 and the 20 task instructions from Table I.

TABLE II
ABLATION STUDY OF TEXT-TO-IMAGE EVALUATION

Front View	Object Resizing	Build Background	CLIP ↑	BLIP ↑	VQA ↑
			0.61168	0.24847	0.84674
✓			0.60904	0.24521	0.86249
	✓		0.61817	0.25967	0.86126
✓	✓		0.61707	0.26068	**0.88541**
✓	✓	✓	**0.61819**	**0.29729**	0.87545

As shown in the Table II, applying Front View estimation resulted in a slight decrease in CLIPScore and BLIPScore, while significantly improving the VQAScore, whereas applying Object Resizing led to consistent improvements across all evaluation metrics. However, when inspecting scenes generated without Front View estimation, we observed failures such as placing a vacuum behind the cabinet instead of next to it, which contradicted the given task instruction. While enabling both configurations slightly lowered the CLIPScore compared to using Object Resizing alone, it improved the BLIPScore and VQAScore without causing any task-related placement failures. The Build Background option, in particular, enhanced realism (see Fig. 4), boosting CLIPScore and BLIPScore. As a result, we conclude that applying both configurations and the option is the most effective strategy for generating semantically aligned and task-aware scenes.

C. Sim-to-Sim Task-Ready Scene Generation for Policy Learning

To evaluate our approach for Question 2, this section evaluates GAIA's capability to produce task-ready scenes for policy learning directly without additional manual configuration.

First, we captured images of simulation environments with a table and obstacles. We then utilized the proposed method to generate a corresponding virtual scene. Subsequently, we provided four task instructions to configure the environment for robot policy learning. As a baseline, we used the RoboGen,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III
AVERAGE SUCCESS RATE (%) OF SIM-TO-SIM POLICY LEARNING
(HIGHER IS BETTER SCORE)

Task Instruction	RoboGen	Ours
Pick up the water bottle on the desk	26.8	**44.6**
Bring me the knife on the cutting board	3.6	**25.2**
Heat up the kettle	23.2	**33.8**
Take out the sauce pot	-	**9.4**

which generates a virtual scene for robot policy learning based solely on a task instruction.

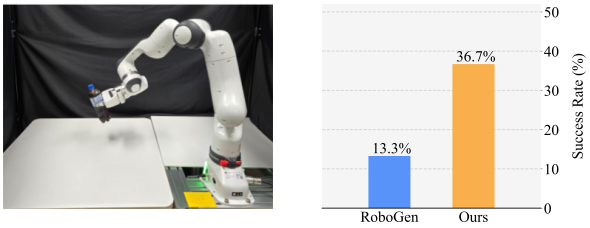
As shown in Fig. 5, we compared the virtual scenes generated for four different task instructions. In the case of RoboGen, it failed to create a scenario requiring inference, such as for the task "*Take out the sauce pot*", which implies the pot should be inside a container. In contrast, our proposed method successfully placed the pot inside a drawer, creating an environment suitable for learning the desired policy.

For each generated virtual scene, we trained a policy using a Soft Actor-Critic (SAC) [51] algorithm. Following RoboGen's policy learning pipeline for rigorous evaluation, the simulation was configured with a suction gripper and the policy's observation space included the target object's pose and orientation, as well as its volumetric information via the minimum and maximum corners of its bounding box. The action space was 7 dimensions. It included a 6D delta pose for the end-effector, which consisted of a 3D delta translation and a 3D delta orientation, as well as a 1D action for opening and closing the gripper. While RoboGen's full pipeline decomposes a task into sequence of subtasks that are solved by various methods such as motion planning, Reinforcement Learning (RL) and trajectory optimization, our evaluation focused on the core policy learned via RL, which corresponds to the main challenge of the original instruction. The reward function for policy learning is generated by a VLM and used directly without any manual modifications. Given the pipeline's sequential structure, policy training for each subtask is initialized from the state where the preceding subtask was completed.

We then evaluated the task success rate of the learned policies over 500 trials per task in the evaluation scene. As presented in Table III, the policies trained in GAIA's scenes demonstrated a higher success rate across all tasks. We have confirmed through this comparison that our method can directly produce task-ready environments well suited for policy learning.

D. Real-to-Sim-to-Real Task-Ready Scene Generation for Transfer Learning

Real-world Policy Transfer. To evaluate sim-to-real transfer for Question 3, we compared two policies. The first was a policy, trained in a GAIA-generated virtual scene on the instruction, "*Pick up the water bottle on the desk*". The second was a policy trained using RoboGen on the identical instruction. For sim-to-real transfer evaluation, we modified the setup to use a 2-finger gripper and trained both policies. Both policies were then deployed on a Franka Research 3 manipulator, which is equipped with this gripper type. Their performance was assessed over 30 trials, each with a randomly sampled bottle position. As shown in Fig. 6, the GAIA-trained policy demonstrated a superior success rate in the real-world execution. We attribute this to GAIA's capability to generate virtual scenes grounded in the spatial information of the input RGB image, which provides more effective preparation for the target task.



(a) (b)

Fig. 6. Real-world policy transfer experiment. (a) Real-world Setup, (b) Success rate of sim-to-real transfer learning.

(a) (b) (c)

Fig. 7. Experiment for the long-horizon task. (a) Real-world scene, (b) GAIA-generated scene, (c) Real robot execution.

Policy Transfer for a Long-Horizon Task. To evalute transferability of the GAIA framework to a long-horizon task, we extended our evaluation to "*Bring me the pencil case from the cabinet*" which involved opening a cabinet and then taking out the pencil case. Following the real-to-sim-to-real pipeline illustrated in Fig. 7, we trained policies in a GAIA-generated virtual scene. These policies executed separate opening and lifting actions sequentially, and success was defined as opening the door and lifting the pencil case by 20 cm. Based on this criterion, the policy achieved a 40% success rate over 10 trials. This result demonstrates that our methodology can be successfully applied to learn complex, long-horizon tasks in a physical environment.

V. CONCLUSIONS

In this paper, we introduce GAIA, a novel methodology for automatically creating interactive, task-ready virtual scenes to train Embodied AI agents and mitigate the sim-to-real gap. Leveraging a pre-trained VLM's reasoning, GAIA processes a single RGB image and a task instruction to retrieve and arrange 3D assets into a task-aware scene. Our experiments confirm GAIA's ability to automate effective simulation setup, evidenced by the real-world success of policies trained within its generated environments. We also acknowledge limitations that present avenues for future work. Our framework's object placement, and its reasoning is constrained by the current VLM and CLIP embeddings. Future work will focus on more precise geometric placement and leveraging improved VLMs to enhance robustness in complex scenarios.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

REFERENCES

[1] Y. S. Narang *et al.*, "Factory: Fast contact for robotic assembly," in *Robotics: Science and Systems*, 2022.

[2] B. Kim and J. Min, "Sim-to-real object pose estimation for random bin picking," in *2024 IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 10 749–10 756.

[3] R. Sarkar *et al.*, "Outfittransformer: Learning outfit representations for fashion recommendation," in *Proceedings of the IEEE/CVF winter Conf. on applications of computer vision*, 2023, pp. 3601–3609.

[4] Z. Dong, X. Liu, B. Chen, P. Polak, and P. Zhang, "Musechat: A conversational music recommendation system for videos," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 12 775–12 785.

[5] V. Ramanujan, T. Nguyen, S. Oh, A. Farhadi, and L. Schmidt, "On the connection between pre-training data diversity and fine-tuning robustness," *Advances in Neural Information Processing Systems*, vol. 36, pp. 66 426–66 437, 2023.

[6] A. Fang *et al.*, "Data determines distributional robustness in contrastive language image pre-training (clip)," in *Int. Conf. on Machine Learning*. PMLR, 2022, pp. 6216–6234.

[7] Y. Xu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation*, vol. 2, no. 4, 2021.

[8] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[9] L. Yang *et al.*, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[10] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, featured Certification.

[11] T. Ren *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[12] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[14] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009.

[15] T. Kim *et al.*, "Realfred: An embodied instruction following benchmark in photo-realistic environments," in *European Conf. on Computer Vision*. Springer, 2024, pp. 346–364.

[16] Y. Mu *et al.*, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 081–25 094, 2023.

[17] F. Xiang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conf. on computer vision and pattern recognition*, 2020, pp. 11 097–11 107.

[18] M. Savva *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF Int. Conf. on computer vision*, 2019, pp. 9339–9347.

[19] Y. Wang *et al.*, "Robogen: Towards unleashing infinite data for automated robot learning via generative simulation," in *Forty-first Int. Conf. on Machine Learning*, 2024.

[20] S. Nasiriany *et al.*, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," in *RSS 2024 Workshop: Data Generation for Robotics*, 2024.

[21] L. Wang *et al.*, "Gensim: Generating robotic simulation tasks via large language models," in *The Twelfth Int. Conf. on Learning Representations*, 2024.

[22] P. Hua *et al.*, "Gensim2: Scaling robot data generation with multi-modal and reasoning llms," in *8th Annual Conf. on Robot Learning*, 2024.

[23] Q. Chen *et al.*, "URDFormer: Constructing interactive realistic scenes from real images via simulation and generative modeling," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*, 2023.

[24] T. Dai *et al.*, "Automated creation of digital cousins for robust policy learning," in *8th Annual Conf. on Robot Learning*, 2024.

[25] J. Tang *et al.*, "Diffuscene: Denoising diffusion models for generative indoor scene synthesis," in *Proceedings of the IEEE/CVF Conf. on computer vision and pattern recognition*, 2024, pp. 20 507–20 518.

[26] Q. A. Wei *et al.*, "Lego-net: Learning regular rearrangements of objects in rooms," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 19 037–19 047.

[27] Y. Yang, B. Jia, P. Zhi, and S. Huang, "Physcene: Physically interactable 3d scene synthesis for embodied ai," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 16 262–16 272.

[28] Y. Yang *et al.*, "Holodeck: Language guided generation of 3d embodied ai environments," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 16 227–16 237.

[29] Y. Wang *et al.*, "Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 575–67 603, 2024.

[30] M. Torne *et al.*, "Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation," *arXiv preprint arXiv:2403.03949*, 2024.

[31] S. Patel *et al.*, "A real-to-sim-to-real approach to robotic manipulation with VLM-generated iterative keypoint rewards," in *2nd CoRL Workshop on Learning Effective Abstractions for Planning*, 2024.

[32] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[33] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[34] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-vocabulary robotic manipulation through mark-based visual prompting," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[35] J. Duan *et al.*, "Manipulate-anything: Automating real-world robots using vision-language models," in *8th Annual Conf. on Robot Learning*, 2024.

[36] Y. J. Ma *et al.*, "Eureka: Human-level reward design via coding large language models," in *The Twelfth Int. Conf. on Learning Representations*, 2024.

[37] M. Zawalski *et al.*, "Robotic control via embodied chain-of-thought reasoning," in *8th Annual Conf. on Robot Learning*, 2024.

[38] J. Duan *et al.*, "AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation," in *The Thirteenth Int. Conf. on Learning Representations*, 2025.

[39] I. Singh *et al.*, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530.

[40] P. V. Georgios Papagiannis, Norman Di Palo and E. Johns, "R+x: Retrieval and execution from everyday human videos," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2025.

[41] H. Yang *et al.*, "Scene synthesis via uncertainty-driven attribute synchronization," in *Proceedings of the IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 5630–5640.

[42] W. Liang *et al.*, "Environment curriculum generation via large language models," in *8th Annual Conf. on Robot Learning*, 2024.

[43] C. Li *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Conf. on Robot Learning*. PMLR, 2023, pp. 80–93.

[44] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. on machine learning*. PmLR, 2021, pp. 8748–8763.

[45] NVIDIA, "Nvidia isaac sim," https://developer.nvidia.com/isaac-sim, 2021, version retrieved 2021.

[46] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[47] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Int. Conf. on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[48] Z. Lin *et al.*, "Evaluating text-to-visual generation with image-to-text generation," in *European Conf. on Computer Vision*. Springer, 2024, pp. 366–384.

[49] A. Shukla, S. Tao, and H. Su, "Maniskill-HAB: A benchmark for low-level manipulation in home rearrangement tasks," in *The Thirteenth Int. Conf. on Learning Representations*, 2025.

[50] J. Cho *et al.*, "Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation," in *The Twelfth Int. Conf. on Learning Representations*, 2024.

[51] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. Conf. on machine learning*. Pmlr, 2018, pp. 1861–1870.