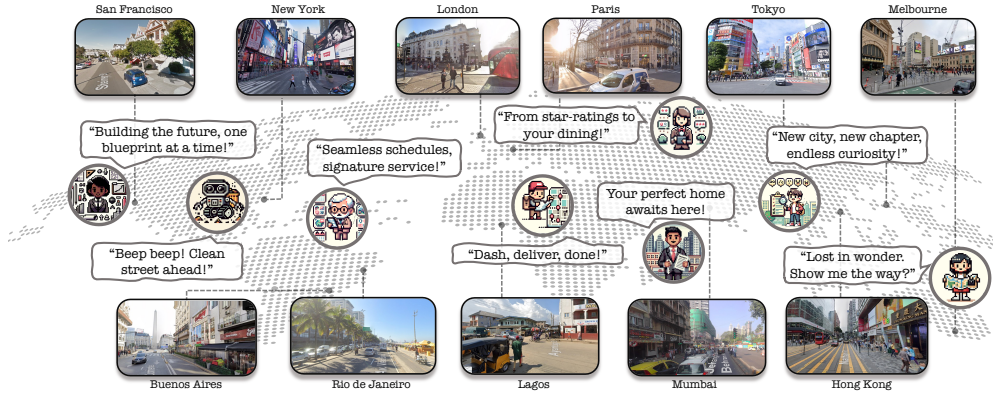


# V-IRL: Grounding Virtual Intelligence in Real Life

Jihan Yang<sup>1\*</sup> Runyu Ding<sup>1</sup> Ellis Brown<sup>2</sup> Xiaojuan Qi<sup>1</sup> Saining Xie<sup>2</sup>

<sup>1</sup>The University of Hong Kong <sup>2</sup>New York University

<https://virl-platform.github.io>



**Fig. 1:** V-IRL agents leverage real-world geospatial information and street view imagery to navigate urban terrains, execute complex tasks, and interact in real-time scenarios. From recommending relevant destinations to assessing city infrastructure to collaboratively giving & following verbal directions—we develop agents that illustrate V-IRL’s current capabilities, flexibility, and utility. Above all else, we present a flexible platform for researchers to harness abundant data from across the globe to create and test diverse autonomous agents.

**Abstract.** There is a sensory gulf between the Earth that humans inhabit and the digital realms in which modern AI agents are created. To develop AI agents that can sense, think, and act as flexibly as humans in real-world settings, it is imperative to bridge the realism gap between the digital and physical worlds. How can we embody agents in an environment as rich and diverse as the one we inhabit, without the constraints imposed by real hardware and control? Towards this end, we introduce V-IRL: a platform that enables agents to scalably interact with the real world in a virtual yet realistic environment. Our platform serves as a playground for developing agents that can accomplish various practical tasks and as a vast testbed for measuring progress in capabilities spanning perception, decision-making, and interaction with real-world data across the entire globe. All V-IRL resources will be open-sourced.

**Keywords:** AI Agents, Embodied AI, Open-world Computer Vision

## 1 Introduction

The advent of large language models (LLMs) has breathed new life into autonomous agent research by offering a universal interface for diverse capabilities,

\* Work conducted during a visit to NYU.

ranging from basic reasoning to complex planning and tool use [72]. While these developments are promising, most of these agents remain confined to text-based environments or simplistic simulations. Visual components in existing agents are either rudimentary—such as simulated tabletop environments [10, 29]—or rely on abstracted representations using ground-truth APIs [27, 68]. Furthermore, the prevalent visual models employed by these agents are trained on photogenic, object-centric Internet images, which fail to capture the unpredictability and diversity of real-world scenes.

This paper aims to bridge this gap between AI agents and the sensory world by grounding them in real-world environments—a crucial step towards developing AI agents that can effectively operate in real-life scenarios. Our novel setting for AI agents *necessitates* rich sensory grounding and perception: virtual embodiment within cities around the globe using real visual and geospatial data.

To this end, we introduce *V-IRL*, a versatile platform for building and testing virtual agents within this novel virtual-real-world setting. *V-IRL* harnesses the power of mapping and street view data, enabling agents to navigate real-world locations, access up-to-date information about their surroundings, and perform practical tasks. With geospatial coordinates at its core, *V-IRL* is flexible and extensible, integrating with arbitrary geospatial platforms and APIs. Moreover, *V-IRL* opens up a vast sea of visual data, allowing a simple and extensible way for researchers to evaluate vision models on realistic data distributions.

We demonstrate the versatility and adaptability of *V-IRL* by developing a series of diverse exemplar agents, each solving a unique and practical task. As these agents hinge upon foundational language and vision models, it is critical to evaluate these models within this setting and their impact on agent performance. We leverage the vast data available through our platform to develop *global scale* benchmarks measuring the performance of underlying vision models on images from diverse geographic and cultural contexts—evaluating their adaptability to shifting environmental, architectural, and language-specific elements. Furthermore, we evaluate the contributions of models to agent performance on challenging tasks. Our results illustrate the potential of *V-IRL* in bridging the gap between virtual agents and visually rich real-world environments, paving the way for future research in this direction.

In summary, our contributions are:

- **V-IRL**: an open-source platform for building and testing agents in a real-world setting that *necessitates* rich sensory grounding and perception—embodiment using real geospatial data and street-view imagery.
- Development of **diverse exemplar agents** that showcase the platform’s versatility and adaptability.
- **Global benchmarks** measuring the performance of foundational language and vision models (1) in isolation using our platform’s real-world data and (2) on end-to-end agent performance in challenging tasks. In addition, we **analyze the robustness of “open-world” vision models to *real-world* data from across the globe.**

We are excited to see how the research community will leverage *V-IRL* to develop and test agents that can understand and interact with the real world.

## 2 Related Work

Here, we ground V-IRL to three streams of research.

**AI Agents.** Agents are autonomous entities capable of perceiving their environment and acting to achieve goals [70]. Historically, agent development has leveraged symbolic and reinforcement learning methods [8, 30, 49], which face issues of scalability and real-world utility. In contrast, the new wave of LLM-driven agents overcomes these challenges with text as a universal interface, enabling natural human interaction and adaptability to various tasks [50, 62, 63, 69, 77]. Moreover, these models equip agents with complex capabilities, such as tool use and collaboration [26, 35, 51, 56, 68, 71, 85]. Yet a critical limitation persists: the agents in this new wave are entirely text-based, devoid of any tangible connection to the visual or sensory aspects of the real world.





**Embodied AI.** Embodied AI studies intelligent agents & robots perceiving and interacting with their environment. A significant challenge in this field is the acquisition of large quantities of realistic data. Consequently, robots are primarily trained in simulated environments [11, 46, 55, 73, 74] to develop skills such as navigation [3, 4, 12] and manipulation [25, 79]. Recent advancements in LLMs [1, 5, 67] have enabled embodied agents to perform long-horizon and open-end tasks in game-engines [27, 29, 39, 45, 60] or human rooms [9, 10, 20, 28, 38]. However, the diversity of tasks and data is still too narrow and simplistic to enable them to operate flexibly in diverse real-world environments.

**Open-World Computer Vision.** Motivated by the success of vision-language models [2, 7, 52, 80] pre-trained on large-scale web-crawled data [15, 32, 57, 61, 66, 75], open-world computer vision has received increasing attention in recent years [18, 19, 23, 33, 34, 37, 48, 65, 76, 83]. However, images and benchmarks sourced from the Internet [6, 17, 22, 31, 33, 54] are unavoidably biased towards specific distributions rather than truly reflecting the real *world* [53]. Because they are trained and evaluated entirely on Internet data, existing “open-world” models are effectively more open-*Internet* than open-*world*.

## 3 Virtual Intelligence in Real Life

To demonstrate the versatility of V-IRL, we use it to instantiate several exemplar agents in our virtual real-world environment. In this section, we engage these agents with tasks that highlight various capabilities of our platform. In Sec. 4, we discuss the technical details of our platform and how it enables agents to interact with the real world.


For illustration, we give V-IRL agents character metadata, including an 8-bit avatar, a name, a short bio, and an intention they are trying to accomplish. More concretely, agents are defined by pipelines that use this character metadata along with our platform’s API and pretrained models to address complex tasks (see Sec. 4). Here we provide a high-level overview of the tasks, highlight the V-IRL capabilities they require, and visualize the agents solving them.

We highlight the specific V-IRL capabilities being employed throughout using tags and corresponding colored underlines:  Map → action,  LLM → reasoning,  Vision → perception, &  Colab → collaboration.

### 3.1 Earthbound Agents

V-*IRL* agents inhabit virtual representations of real cities around the globe. At the core of this representation are *geographic coordinates* corresponding to points on the Earth’s surface. Using these coordinates, V-*IRL* allows virtual agents to *ground* themselves in the real world using maps, street view imagery, information about nearby destinations, and additional data from arbitrary geospatial APIs.

Route Optimizer
ENV Map



**Name:** Peng    **Age:** 21    **Loc:** NYC

**Bio:** Originally from Chengdu, Sichuan, Peng is a student at PKU. He just arrived for a semester abroad at NYC, and is couch surfing until he gets settled.

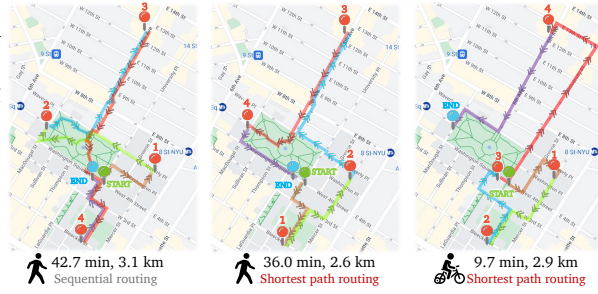
**Intention:** Peng needs to visit five locations around the city: his University Card Center, Residence Hall, Research Center, Library, and Student Center.

**Task:** Given a starting address and a list of waypoints, plan the shortest route to all waypoints and then follow it on street view.

**Takeaway:** V-*IRL* instantiates agents with *real* geospatial information, and enables useful tasks like route optimization.

*Peng needs to visit several locations throughout the city to get documents signed for registration as a visiting student. . .*

By leveraging geolocation and mapping capabilities within our V-*IRL* environment, Peng saves 7 minutes by walking along the shortest path as opposed to in-order waypoint visitation. An illustration is shown in Fig. 2.



**Fig. 2:** Optimizing Peng’s travel route to five places.

### 3.2 Language-Driven Agents

To tackle more complex tasks, we follow the pattern of language-driven agents [72]. LLMs enable agents to flexibly reason, plan, and use external tools & APIs.

Place Recommender
ENV Map LM LLM



**Name:** Aria    **Age:** 26    **Loc:** NYC

**Bio:** A 3rd year graduate student who loves to try new restaurants. She is always looking for new places to try, and shares her favorite spots on her blog!

**Intention:** Pick out a lunch spot that Peng might like.



**Name:** Vivek    **Age:** 35    **Loc:** NYC

**Bio:** A tech-savvy estate agent who combines his local knowledge with online tools like Zillow to find the perfect homes for his clients in the bustling city.

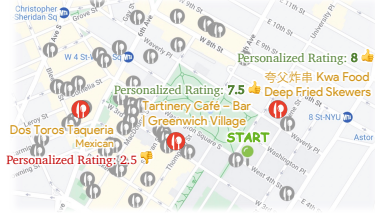
**Intention:** Help Peng find a place to live for the semester.

**Task:** Given specific location, background, and intention, synthesize reviews of nearby businesses to provide a recommendation.

**Takeaway:** V-*IRL* exposes rich real-world information to agents that they can use for real-world tasks.

*Peng is starving for some lunch but doesn’t know where to eat. . . Luckily, he met a nice grad student Aria during his errands who can help him find a good spot. . .*





Aria searches for possible restaurants nearby. She then synthesizes public reviews to make final recommendations via GPT-4. As Peng is new to the city and originally from Sichuan, she recommends a spicy Chinese joint *Kwa Food Deep Fried Skewers* to give him a taste of home.

Peng hires Vivek to find an apartment in East Village, Jersey City, or Long Island City for \$1k–\$3k monthly, close to a gym, supermarket, and public transit. . .

Recommendations	Rental Information
<p><b>Personalized rating: 7.5/10</b> 🌟</p> <p>The apartment is well-located with easy access to supermarkets, public transport, and a gym, which aligns with Peng's requirements. However, the price may not be cost-effective for a student.</p>	<p>"address": "12E, New York, NY 11101",  "rent": \$2904, "type": "Apartment",  "sqr": 450, "bedrooms": 0,  "bathrooms": 1,</p>
<p><b>Personalized rating: 8/10</b> 🌟</p> <p>The apartment is well-located near a supermarket and gym, which aligns with Peng's lifestyle. Multiple bus stations are nearby, but the lack of a close subway station may affect his commute.</p>	<p>"address": "508, New York, NY 11101",  "rent": \$1986, "type": "Apartment",  "sqr": 800, "bedrooms": 1,  "bathrooms": 1,</p>
<p><b>Personalized rating: 2/10</b> 🌟</p> <p>The estate lacks nearby supermarkets, bus, subway stations, and gyms, which are essential for Peng's requirements.</p>	<p>"address": "871, Jersey City, NJ 07302", "rent": \$711, "type": "Apartment", "sqr": 1, "year built": 1992,</p>

Vivek uses real estate APIs to find potential apartments in Peng's desired regions and price range. For each candidate, he researches its proximity to the places Peng cares about. Synthesizing these factors, Vivek provides a holistic rating with reasoning using GPT-4. His top recommendation is a cost-effective 1 bedroom apartment for \$1986/mo, which is close to a supermarket, 2 bus stations, and a gym.

### 3.3 Visually Grounded Agents

Although language-driven agents can address some real-world tasks using external tools, their reliance solely on text-based information limits their applicability to tasks where *visual grounding* is required. In contrast, *real sensory input* is integral to many daily human activities—allowing a deep connection to and understanding of the world around us. Agents can leverage street view imagery through the V-IRL platform to *visually ground* themselves in the real world—opening up a wide range of *perception-driven tasks*.

Urban Planner

Map
 Vision

**Name:** Imani    **Age:** 42    **Loc:** NYC

**Bio:** A sustainable urban development graduate, Imani is passionate about maintaining a harmonious balance between nature and urban ecosystems.

**Intention:** Use RX-399 to collect first-person data for her studies.

**Name:** RX-399    **Age:** Unk.    **Loc:** NYC

**Bio:** RX-399 is a robot with advanced detection, localization, and navigation systems.

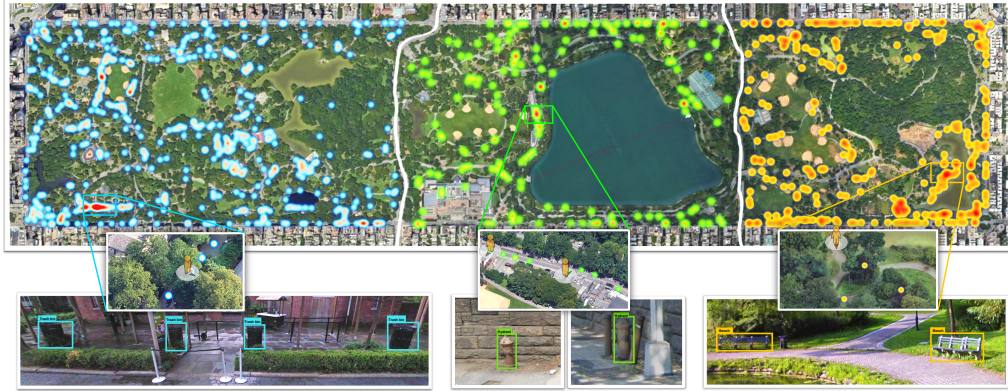
**Intention:** Localize and count pre-defined categories to the user in specified regions.

**Task:** Record the location of all instances of any specified objects (e.g., trash bins, hydrants, benches) in a specified region.

**Takeaway:** V-IRL enables realistic open-world applications requiring vast geospatial and first-person visual information.

Imani is analyzing the distribution of trash bins, fire hydrants, and park benches in Central Park, NYC for a project with the Parks and Recreation Department. . .

Imani sets routes spanning Central Park and objects of interest for RX-399, who traverses the routes and records all detected instances. After RX-399 finishes its route, Imani analyzes the collected data at different levels of detail. As depicted in Fig. 3, the coarsest level shows general distributions of trash bins, hydrants, and benches in the park. Imani can also zoom in to specific regions, where lighter



**Fig. 3:** Imani's visualization of trash bins, fire hydrants, & park benches in NYC's Central Park using data collected by RX-399.

colors represent positions with more unique instances identified. The following table presents RX-399's counting report:

Category	Trash Bin	Fire Hydrant	Park Bench*
Count	1059	727	1015

(\*Note: contiguous benches counted as one instance). By retrieving geotagged sensory-rich data within RX-399, Imani can also inspect the detection results for each object to verify the reliability of RX-399's reports.

During RX-399's traversal, it can avoid double-counting previously seen objects by using feature matching to check for duplicates across different viewpoints (see figure illustration on the right).



**Intentional Explorer** ENV Map L LLM Vision

**Name:** Hiro    **Age:** 22    **Loc:** HK

**Bio:** A seasoned traveler, Hiro thrives in unknown territories. He enjoys getting lost in new places instead of following the travel guide.

**Intention:** Hiro is looking for an authentic lunch spot that is not too spicy.

**Task:** Explore on foot (in street view) looking for a destination that fulfills a certain intention (e.g., lunch, shopping, etc.)

**Takeaway:** Agents can utilize visual detectors, VLMs and LLMs to iteratively perceive, decide, and interact in the environment.

*Hiro starts a new journey in Hong Kong. He decides to explore without a specific destination, looking for a good local lunch spot with food that's not too spicy...*

As depicted in Fig. 4, starting at 1, Hiro walks down the street and encounters the first intersection. Thanks to the interactive and sensory-rich environment, he can adjust his pose to fetch street views for each possible path. Using VQA on these views, he decides to turn left:


★ *Residential buildings on the left road indicate cozy and family-run local food... A better choice than the others!*

Then, after exploring for a block, he encounters the second intersection where he looks around and decides to turn right:

★ *Looks like there are some local food spots this way...*



Fig. 4: Hiro’s lunch exploration procedure in HK.





Finally, at the end of the block , Hiro discovers another lunch spot called “Xintianfa”. He decides to dine there after reading numerous online reviews praising its authentic cuisine and diverse menu. See the low-level case study for technical details behind Hiro in Appendix F.


### 3.4 Collaborative Agents

Humans work together to solve complex real-world tasks. This collaboration promotes efficiency and effectiveness by decomposing a complex task into simpler sub-tasks, allowing each to be handled by an expert in its domain. Grounded in the world via our platform, V-IRL agents can leverage geospatial data and street view imagery to collaborate with other agents as well as with human users.

**Agent-Agent Collaboration** As with previous agents, collaborative agents are designed for specific tasks; however, they can handle objectives beyond their expertise through collaboration with each other.

Tourist

 Map
  LLM
  Vision
  Colab



**Name:** Ling    **Age:** 25    **Loc:** NYC/SF/HK

**Bio:** Ling is a spirited traveler from Taipei who is always eager to explore new cities and cultures. She is unafraid of asking locals for help when she's lost!

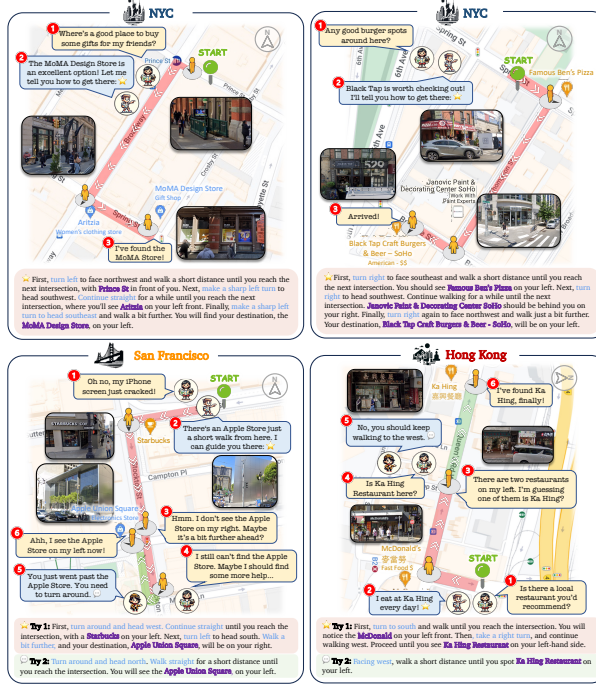
**Intention:** NYC: find gifts for friends back home; go to a famous restaurant. SF: find a store to repair a broken iPhone. HK: try some authentic local food.

**Task:** (i) Ask for directions to a specific location from a nearby Local agent, who previews the route with the map and streetview and then gives directions in natural language, mentioning major intersections and landmarks. (ii) Follow these directions in streetview, and if lost, ask another Local agent for assistance.

**Takeaway:** Agents can collaborate to solve complex tasks that are beyond their individual expertise.

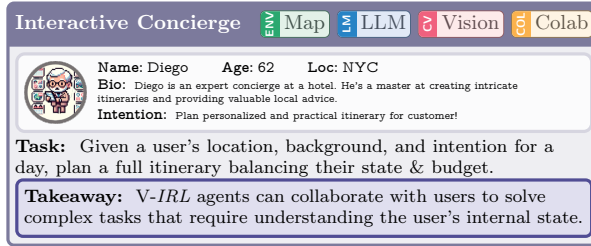
*Ling travels to cities around the world. She seeks out authentic experiences and is always unafraid to ask for help from Locals whenever she finds herself lost. . .*

After obtaining route descriptions from Locals, Ling starts her journey—as shown in Fig. 5. Grounded in our embodied platform, Ling can adjust her pose and identify visual landmarks along the streets using open-world recognition and her map. Correctly recognizing these landmarks helps her to make correct decisions about where to change direction, move forward, and stop, as seen in



**Fig. 5:** Ling’s collaborations with Local. Red and green for the first and second attempts, respectively.

**Human-Agent Collaboration** Grounded in the same environment we humans inhabit, V-*IRL* agents can collaborate with and assist real human users.



*As a university student in NYC, you are excited to spend a day exploring lesser-known and tranquil places. Your friend recommended Diego, who is known for his professionalism in planning practical and personalized itineraries.*

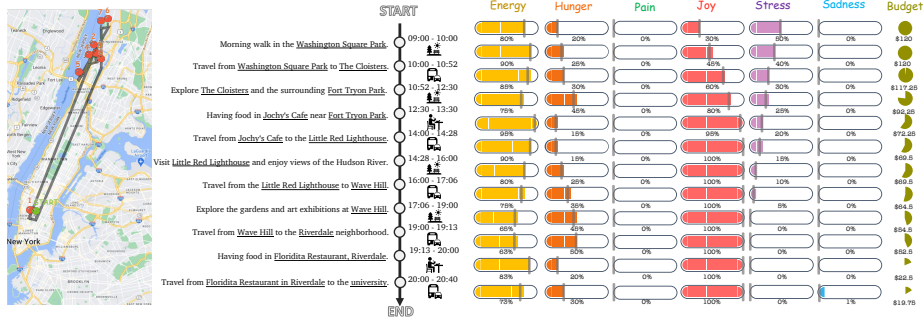
As depicted in Fig. 6, Diego’s itinerary is tailored to *your* needs. Diego not only considers your physical and mental interoceptive status, budget for each activity, but also anticipates your status changes and cost when you follow each event. He can take into account *real travel times* from the V-*IRL* platform and select suitable destinations by *collaborating* with another recommendation agent. For comparison, see the “ungrounded” LLM-only concierge in Appendix B.2.

Also, as shown in Fig. 7, you can intervene in Diego’s planning process by adjusting your interoceptive status or by providing verbal feedback. In response,

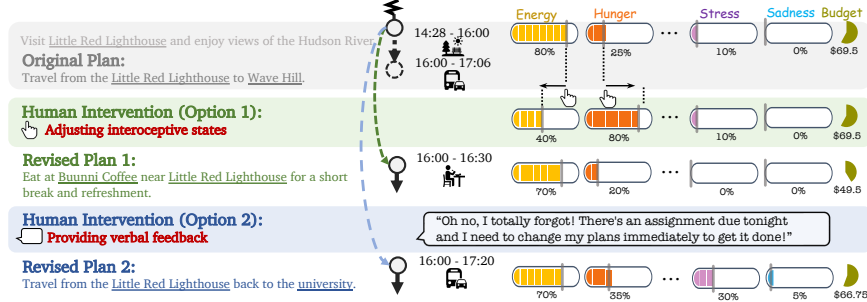
the top two NYC cases in Fig. 5. The success of these decisions made by Ling with *GPT-4* relies on the real-sensory input for visual grounding and the interactive environment from V-*IRL*.

Nevertheless, Ling may occasionally fail to find the destination. In the SF example of Fig. 5, Ling passes by the Apple Store because only its stainless steel wall is visible from her viewpoint. In the HK case, Ling mistakes another restaurant for her destination and stops prematurely. Fortunately, when she makes these mistakes, Ling can ask another Local agent for another round of navigation until it eventually leads her to the destination.





**Fig. 6: The Perfect Day Itinerary:** Crafted by Diego, our iterative concierge agent, this schedule is meticulously tailored, accounting for your mental and physical well-being and budget variations as your day unfolds.



**Fig. 7: Diego adapts original plan to suit user's intervention.**

Diego promptly revises his original plan to accommodate your demands, and re-estimates your state changes after his revision.

Finally, using V-IRL's street views and map, Diego can traverse regions of interest scouting for potential scenic viewpoints for you to visit as shown in Fig. 8. He uses VQA to rate and assess each captured view, and adds the highest-rated locations to your itinerary.



**Fig. 8: Diego traverses regions of interest to find scenic locations for your itinerary.**

## 4 System Fundamentals

This section introduces our system's core: a platform designed for perception-driven agents that transforms real-world cities around the world into a vast virtual playground where agents can be constructed to solve practical tasks. At its heart, V-IRL is comprised of a hierarchical architecture (see Fig. 9). The *platform* lies at the foundation—providing the underlying components and infrastructure for agents to employ. Higher level *capabilities* of **Perception**, **Reasoning**, **Action**, and **Collaboration** emerge from the platform's components. Finally, *agents* leverage these capabilities and user-defined metadata in task-specific routines to solve tasks.

#### 4.1 Agent Definition

In our system, agent behavior is shaped by user-defined metadata, including a background, an intended goal, and an interoceptive state. The *background* provides the context necessary to instantiate the agent in the real world (location), and to guide its reasoning and decision-making (biography). *Intentions* outline agents’ purpose within the environment. An agent’s *interoceptive state* reflects its internal mental and physical status—varying over time and influencing its behavior. This novel concept is crucial to AI agents for enhancing collaboration with humans (see Sec. 3.4). Concretely, agents are developed by writing task-specific `run()` routines that leverage the various components of our platform and the agent’s metadata to solve tasks.

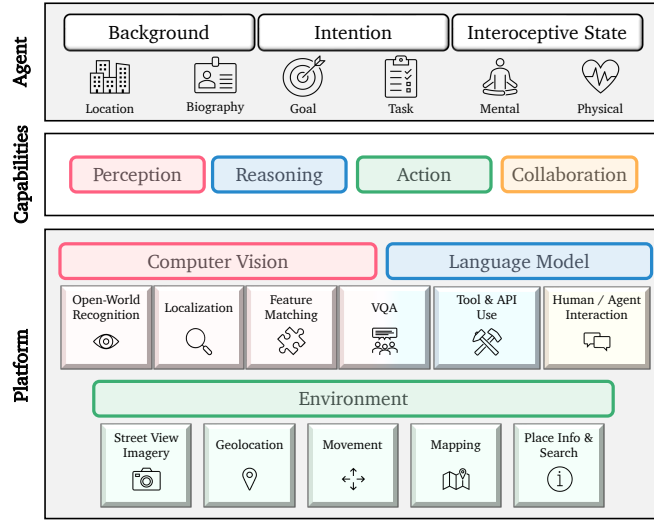


Fig. 9: Hierarchical V-IRL architecture described in Sec. 4.

#### 4.2 Platform Components



Next, we delve into platform components, which provide the infrastructure to instantiate capabilities, execute agent actions, and ground agents in real world.

**Environment (Action)** ENV Environment components are responsible for grounding agents in the world around them: providing a navigable representation of real cities (see Sec. 3.1). Geographic coordinates serve as the link between the world and our virtual representation of it. Leveraging the Google Maps Platform (GMP) [24], V-IRL enables agents to access street view imagery, query valid movements, retrieve information about nearby locations, and plan routes. As these coordinates and location information are bound to the real world, they also provide a natural interface with external tools that leverage geolocation—such as real estate APIs (see Sec. 3.2). More technical details in Appendix D.

**Vision (Perception)** P Perception components enable agents to process the sensory-rich data provided by the ENV Environment, especially street view imagery. Pretrained localization models [37] give agents a precise spatial understanding of their environment. This allows RX-399 to identify and count instances of objects, and Hiro to pick out specific businesses to look up with the ENV Environment (Sec. 3.3). While



localization models allow for precise interaction with perceptive input, open-world recognition models [52] are more general, and allow agents to detect a wider range of objects in their field of view (*e.g.*, Tourist searches for the Apple Store). Pretrained feature matching models [40] provide an understanding of continuity across views of the same location, and enable agents to identify & deduplicate instances of the same object from different viewpoints (Sec. 3.3). Multimodal models with VQA & Captioning capabilities [36] bridge the perceptual world with natural language, and are essential for integration with reasoning (Sec. 3.3).

**Language (Reasoning & Collaboration)**  Reasoning components allow decision making based on information from perception and the environment. LLMs such as GPT-4 [1] and Llama 2 [67] interface across various APIs (Sec. 3.2), transforming environmental data and perceptual outputs into actionable insights. They also enable  Collaboration between agents or with humans through natural language (Sec. 3.4). See case studies in Appendix E for details.

### 4.3 V-IRL Capabilities

Our platform’s components can be flexibly combined to exhibit a vast array of capabilities. In Sec. 3, we present agents that exhibit increasingly complex behaviors, each requiring more components of the platform. From simple combinations, like the Route Optimizer (Sec. 3.1), to more complex arrangements, like the Tourist (Sec. 3.4), our system showcases the versatility and potential of the V-IRL platform to be applied to various real-world scenarios. To facilitate understanding, we perform both high-level and low-level case studies of how V-IRL’s components are combined to create complex V-IRL agents “Diego” and “Hiro” in Appendix E and F, respectively.

## 5 V-IRL Benchmarks

In the previous sections, we illustrate the primary benefit of the V-IRL platform: seamless access to first-person street-view imagery and descriptive information about real-world cities across the globe. This *scalable* source of *truly open-world* data can be harnessed to test core component models and agent capabilities. We propose three V-IRL benchmarks: two evaluating vision-language models on open-world vision tasks (Secs. 5.2 and 5.3), and one evaluating end-to-end agent performance (Sec. 5.4). More benchmark details and results are in Appendix G.

### 5.1 Automated Data and Annotation Collection

To allow our V-IRL benchmarks to scale globally, we develop an automatic data/annotation construction pipeline instead of crawling and manually annotating limited data. This allows models to be conveniently tested worldwide, provided there is access to Google Street Views [24].

**Region Selection.** Though our benchmark is feasible across all regions covered by the GMP, we select 14 districts across 12 cities from 6 continents to ensure coverage of a diverse data distribution while keeping inference costs affordable. The detailed locations of these regions are listed in Appendix G.1.

**Vision and Place Data Collection.** Within each region, we collect geolocations with available street views, place information, and place-centric images.

## 5.2 V-IRL Place: Detection

Every day, humans traverse cities, moving between varied places to fulfill a range of goals, like the Intentional Explorer agent (Sec. 3.3). We assess the performance of vision models on the everyday human activity of *localizing places* using street view imagery and associated place data.

**Setups.** We modify RX-399 (Sec. 3.3) to traverse 28 polygonal areas from the 14 districts while localizing & identifying 20 types of places.

**Benchmarked Models.** We evaluate five open-world detection models: GroundingDINO [43], GLIP [37], Owl-ViT [48], OpenSeeD [82] and Owl-ViT v2 [47]. We also implement a straightforward baseline, CLIP (w/ GLIP proposal), which involves reclassifying the categories of GLIP proposals with CLIP [52].

**Evaluation.** We evaluate the models based on localization recall, which is quantified as  $\frac{N_{tp}}{N_{tp}+N_{fn}}$ , where  $N_{tp}$  and  $N_{fn}$  represents the number of correctly localized places and missed places, respectively. See more details in Appendix G.2.

**Results.** Tab. 1 shows that open-world detectors like GroundingDINO [43], Owl-ViT [48] and GLIP [37] are biased towards certain place types such as **school**, **cafe**, and **park**, respectively. In contrast, CLIP (w/ GLIP proposal) can identify a broader spectrum of place types. This is mainly caused by the category bias in object detection datasets with a limited vocabulary. Hence, even if detectors like Owl-ViT are initialized with CLIP, their vocabulary space narrows down due to fine-tuning. These results suggest that cascading category-agnostic object proposals to zero-shot recognizers appears promising for “real” open-world detection—especially for less common categories in object detection datasets. See full results and more analysis in G.2.

**Table 1:** Benchmark results on V-IRL Place Detection.  $AR^{10}$  and  $AR^{20}$  denote average recall on subsampled 10 and all 20 place categories, respectively.

Place Types															$AR^{10}$	$AR^{20}$
GroundingDINO [43]	0.0	0.0	0.0	0.0	0.0	4.9	0.0	0.0	100.0	0.0	11.7	5.8				
Owl-ViT [48]	0.0	61.0	0.0	0.0	0.0	2.4	0.3	0.0	0.0	0.0	7.1	7.1				
GLIP [37]	20.0	0.0	100.0	0.0	0.0	0.0	18.4	0.0	0.0	0.0	15.4	9.0				
OpenSeeD [82]	60.0	11.9	50.0	0.0	0.0	0.0	20.5	0.0	0.0	16.7	17.7	16.7				
Owl-ViT v2 [47]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.4				
CLIP [52] (w/ GLIP proposal)	60.0	6.8	50.0	40.0	25.0	29.3	14.7	0.0	0.0	16.7	26.9	23.7				

## 5.3 V-IRL Place: Recognition and VQA

In contrast to the challenging V-IRL place detection task using street view imagery alone, in real life, humans can recognize businesses by taking a closer, place-centric look. We assess existing vision models in this manner on two perception tasks based on place-centric images: *i*) recognizing specific place types; *ii*) identifying human intentions via Vision Question Answering (VQA).

**Setups.** For recognition, we assess 10 open-world recognition models on identifying a place’s type from 96 options using place-centric images. For VQA, we evaluate 8 multi-modal large language models (MM-LLM) to determine viable human intentions from a four-option multiple-choice (see more in Appendix G.3).

**Evaluation.** We adopt mean accuracy (mAcc) to evaluate both place recognition and VQA tasks. For place VQA, we follow MMBench [44] to conduct circular evaluation and GPT-assisted answer parsing.

**Results.** Tab. 2 shows that CLIP (L/14@336px) outperforms even the biggest version of Eva-02-CLIP and SigLIP in the V-IRL recognition task, highlighting the high-quality data used to train CLIP [52]. The bottom of the table shows that BLIP2 [36], InstructBLIP [16], and LLaVA-1.5 [41] excel at intention VQA, whereas others struggle. We note that these three top-performing MM-LLMs provide consistent answers in the circular evaluation, while others frequently fail due to inconsistent selections. Moreover, vision models perform better on intention VQA over place-type recognition, suggesting direct prompts about human intention could be more effective for intention-driven tasks. We provide more results and analysis in Appendix G.3.

**Table 2:** Benchmark results on V-IRL Place recognition and VQA. Green for increased resolution models; Blue denotes model parameter scaling.

	Model	#Param	mAcc (%)
<b>V-IRL Place Recognition</b>			
CLIP [52]	ViT-B/32	151M	18.2
CLIP [52]	ViT-L/14	428M	37.2
CLIP [52]	ViT-L/14@336px	428M	41.3
OpenCLIP [15]	ViT-B/32	151M	21.2
OpenCLIP [15]	ViT-L/14	428M	31.0
Eva-02-CLIP [64]	ViT-B/16	150M	19.5
Eva-02-CLIP [64]	ViT-L/14	428M	34.2
Eva-02-CLIP [64]	ViT-L/14@336px	428M	40.7
SigLIP [81]	ViT-B/16	203M	29.5
SigLIP [81]	ViT-L/16@384px	652M	37.3
<b>V-IRL Place VQA</b>			
MiniGPT-4 [84]	Vicuna-13B-v0	14.0B	3.9
mPLUG-Owl [78]	LLaMA-7B	7.2B	5.5
Shikra [14]	Vicuna-7B	7.2B	10.9
BLIP-2 [36]	FlanT5 <sub>XXL</sub>	12.1B	69.6
InstructBLIP [16]	FlanT5 <sub>XXL</sub>	12.0B	68.0
LLaVA [42]	Vicuna-13B-v1.3	13.4B	23.5
LLaVA-1.5 [41]	Vicuna-7B-v1.5	7.2B	60.1
LLaVA-1.5 [41]	Vicuna-13B-v1.5	13.4B	61.9

#### 5.4 V-IRL Vision-Language Navigation

As discussed in Sec. 3.3, Intentional Explorer and Tourist agents require coordination between vision models and language models to accomplish vision-language tasks. To investigate the effect of various models on end-to-end agent performance, we develop an embodied task that jointly tests vision and language models: Vision-Language Navigation (VLN). In VLN, agents navigate to a desired destination by following textual directions using only raw street views.

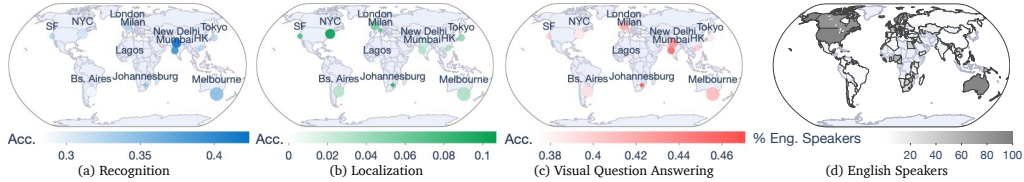
**Setup.** We adopt the Tourist implementation from Sec. 3.4 and swap its recognition component with the various benchmarked models. These models are used to identify visual landmarks during navigation. Subsequently, GPT-4 [1] predicts the next action according to the recognition results. Navigation instructions are generated using the Local agent. Recent work VELMA [59] attempts to enhance VLN by leveraging LLMs on existing datasets [13, 58]. In contrast, our V-IRL VLN benchmark evaluates vision models and their coordination with language models across a global data scale. See more details in Appendix G.4.

**Benchmarked methods.** Four approaches are evaluated to recognize landmarks during navigation: (i) Oracle that searches nearby landmarks with GMP [24]; (ii) Zero-shot recognizers CLIP [52] & EVA-CLIP [64]; (iii) Multi-modal LLM LLaVA-1.5 [41]; (iv) An OCR model [21] to extract text in street views followed by GPT answer parsing. Implementation details are provided in Appendix G.4.

**Evaluation.** We primarily measure navigation success rate (*Success*), defining success as the navigator stopping within 25 meters of the destination. In ad-

**Table 3:** Results on V-*IRL* VLN miniset. We test various CLIP-based models, MM LLM, and OCR model with GPT postprocessing.

Method	Success	Start Intersection		Stop	
		Reac	Arr	Reac	Arr
Oracle (No Vision)	1.0	1.0	1.0	1.0	1.0
CLIP (B/32) [52]	0.22	1.0	0.86	0.84	0.83
CLIP (L/14@336px) [52]	0.44	0.83	0.73	0.94	0.67
EVA-02-CLIP (BigE/14-plus) [64]	0.39	0.89	0.77	0.94	0.72
EVA-02-CLIP (L/14@336px) [64]	0.22	1.0	0.82	0.83	0.78
LLaVA-1.5-13B [41]	0.11	0.61	0.55	1.0	0.56
PP-OCR [21] (+ GPT3.5)	0.28	0.89	0.73	0.94	0.72

**Fig. 10:** City-level visualization of V-*IRL* benchmark results.

dition, as navigation success is mainly influenced by the agent’s actions at key positions (*i.e.*, start positions, intersections and stop positions), we also evaluate the arrival ratio (*Arr*) and reaction accuracy (*Reac*) for each route. *Arr* denotes the percentage of key positions reached, while *Reac* measures the accuracy of the agent’s action predictions at these key positions. To save GPT-4 resources, we mainly compare vision modules on a 10% mini-set comprising 18 routes from 9 regions. See Appendix G.4 for full-set results with CLIP and Oracle.

**Results.** Table 3 shows that, with oracle landmark information, powerful LLMs can impressively comprehend navigation instructions and thus make accurate decisions. However, when relying on vision models to identify landmarks in street views, the success rate drops dramatically—suggesting that the perception of vision models is noisy and misguides LLMs’ decision-making. Among these recognizers, larger variants of CLIP [52] and EVA-02-CLIP [64] perform better, highlighting the benefits of model scaling. LLaVA-1.5 [41] shows inferior results with CLIP (L/14@336px) as its vision encoder, possibly due to the alignment tax [1] from instruction tuning. Further, PP-OCR [21] (+ GPT-3.5) achieves a 28% success rate, emphasizing the value of OCR in visual landmark recognition.

### 5.5 Geographic Diversity

Spanning 12 cities across the globe, our V-*IRL* benchmarks provide an opportunity to analyze the inherent model biases across different regions. As depicted in Fig. 10, vision models demonstrate subpar performance on all three benchmark tasks in Lagos, Tokyo, Hong Kong, and Buenos Aires. Vision models might struggle in Lagos due to its non-traditional street views relative to more developed cities (see street views in Fig. 1). For cities like Tokyo, Hong Kong, and Buenos Aires, an intriguing observation is their primary use of non-English languages in street views (Fig. 10 (d) <sup>1</sup> and Fig. 1). This suggests that existing vision models may face challenges when deployed in non-English-dominant countries.

<sup>1</sup> Source: [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_English-speaking\\_population](https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population)

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023) [3](#), [11](#), [13](#), [14](#)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) [3](#)
3. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018) [3](#)
4. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018) [3](#)
5. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: PaLM 2 technical report. arXiv preprint arXiv:2305.10403 (2023) [3](#)
6. Asano, Y.M., Rupprecht, C., Zisserman, A., Vedaldi, A.: PASS: An imagenet replacement for self-supervised pretraining without humans. In: NeurIPS (2021) [3](#)
7. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: OpenFlamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023) [3](#)
8. Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al.: Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680 (2019) [3](#)
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.W.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023) [3](#)
10. Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., et al.: Do As I Can, Not As I Say: Grounding language in robotic affordances. In: CoRL (2023) [2](#), [3](#)
11. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from rgb-d data in indoor environments. In: 3DV (2017) [3](#)
12. Chaplot, D.S., Gandhi, D.P., Gupta, A., Salakhutdinov, R.R.: Object goal navigation using goal-oriented semantic exploration. In: NeurIPS (2020) [3](#)
13. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. In: CVPR (2019) [13](#)
14. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) [13](#)

15. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023) 3, 13
16. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In: NeurIPS (2023) 13
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009) 3
18. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7010–7019 (2023) 3
19. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 3
20. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-E: An Embodied Multimodal Language Model. In: ICML (2023) 3
21. Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., et al.: PP-OCR: A practical ultra lightweight ocr system. arxiv 2020. arXiv preprint arXiv:2009.09941 (2020) 13, 14
22. Dubey, A., Ramanathan, V., Pentland, A., Mahajan, D.: Adaptive methods for real-world domain generalization. In: CVPR (2021) 3
23. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV (2022) 3
24. Google Map Team: Google Map Platform. <https://mapsplatform.google.com/> 10, 11, 13
25. Gu, S., Holly, E., Lillicrap, T., Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: ICRA (2017) 3
26. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., Schmidhuber, J.: MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In: ICLR (2023) 3
27. Huang, W., Abbeel, P., Pathak, D., Mordatch, I.: Language Models As Zero-Shot Planners: Extracting actionable knowledge for embodied agents. In: ICML (2022) 2, 3
28. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: VoxPoser: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973 (2023) 3
29. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al.: Inner Monologue: Embodied reasoning through planning with language models. In: CoRL (2022) 2, 3
30. Küttler, H., Nardelli, N., Miller, A., Raileanu, R., Selvatici, M., Grefenstette, E., Rocktäschel, T.: The nethack learning environment. In: NeurIPS (2020) 3
31. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020) 3
32. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., et al.: OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In: NeurIPS (2023) 3



33. Li, A.C., Brown, E., Efros, A.A., Pathak, D.: Internet explorer: Targeted representation learning on the open web. In: International Conference on Machine Learning. pp. 19385–19406. PMLR (2023) [3](#)
34. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022) [3](#)
35. Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B.: CAMEL: Communicative agents for "mind" exploration of large language model society. In: NeurIPS (2023) [3](#)
36. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) [11](#), [13](#)
37. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR (2022) [3](#), [10](#), [12](#)
38. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as Policies: Language model programs for embodied control. In: ICRA (2023) [3](#)
39. Lin, K., Agia, C., Migimatsu, T., Pavone, M., Bohg, J.: Text2Motion: From natural language instructions to feasible plans. arXiv preprint arXiv:2303.12153 (2023) [3](#)
40. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023) [11](#)
41. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv:2310.03744 (2023) [13](#), [14](#)
42. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) [13](#)
43. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) [12](#)
44. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: MMBench: Is Your Multi-modal Model an All-around Player? arXiv preprint arXiv:2307.06281 (2023) [13](#)
45. Liu, Z., Bahety, A., Song, S.: REFLECT: Summarizing robot experiences for failure explanation and correction. In: CoRL (2023) [3](#)
46. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac Gym: High performance gpu-based physics simulation for robot learning. In: NeurIPS (2021) [3](#)
47. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024) [12](#)
48. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple Open-Vocabulary Object Detection with Vision Transformers. In: ECCV (2022) [3](#), [12](#)
49. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013) [3](#)
50. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332 (2021) [3](#)
51. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative Agents: Interactive simulacra of human behavior. In: UIST (2023) [3](#)

52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 3, 11, 12, 13, 14
53. Ramaswamy, V.V., Lin, S.Y., Zhao, D., Adcock, A.B., van der Maaten, L., Ghadiyaram, D., Russakovsky, O.: GeoDE: a geographically diverse evaluation dataset for object recognition. In: NeurIPS (2023) 3
54. Rojas, W.A.G., Diamos, S., Kini, K.R., Kanter, D., Reddi, V.J., Coleman, C.: The Dollar Street Dataset: Images representing the geographic and socioeconomic diversity of the world. In: NeurIPS (2022) 3
55. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV (2019) 3
56. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. In: NeurIPS (2023) 3
57. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022) 3
58. Schumann, R., Riezler, S.: Generating landmark navigation instructions from maps as a graph-to-text problem. In: ACL (2020) 13
59. Schumann, R., Zhu, W., Feng, W., Fu, T.J., Riezler, S., Wang, W.Y.: VELMA: Verbalization embodiment of llm agents for vision and language navigation in street view. arXiv preprint arXiv:2307.06082 (2023) 13
60. Shao, H., Hu, Y., Wang, L., Waslander, S.L., Liu, Y., Li, H.: LMDrive: Closed-loop end-to-end driving with large language models. arXiv preprint arXiv:2312.07488 (2023) 3
61. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 3
62. Shinn, N., Labash, B., Gopinath, A.: Reflexion: an autonomous agent with dynamic memory and self-reflection. In: NeurIPS (2023) 3
63. Significant Gravitas: AutoGPT. <https://github.com/Significant-Gravitas/AutoGPT> (2023) 3
64. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv preprint arXiv:2303.15389 (2023) 13, 14
65. Tang, L., Tian, Z., Li, K., He, C., Zhou, H., Zhao, H., Li, X., Jia, J.: Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. arXiv preprint arXiv:2407.05342 (2024) 3
66. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. Communications of the ACM (2016) 3
67. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: LLAMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 3, 11
68. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023) 2, 3
69. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022) 3

70. Wooldridge, M., Jennings, N.R.: Intelligent Agents: Theory and practice. The knowledge engineering review (1995) 3
71. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C.: AutoGen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155 (2023) 3
72. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023) 2, 4
73. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson Env: Real-world perception for embodied agents. In: CVPR (2018) 3
74. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: SAPIEN: A simulated part-based interactive environment. In: CVPR (2020) 3
75. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023) 3
76. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19823–19832 (2024) 3
77. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In: ICLR (2023) 3
78. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mPLUG-Owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 13
79. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., et al.: HomeRobot: Open-vocabulary mobile manipulation. In: CoRL (2023) 3
80. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021) 3
81. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. arXiv preprint arXiv:2303.15343 (2023) 13
82. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023) 12
83. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022) 3
84. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv:2304.10592 (2023) 13
85. Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., Qiao, Y., Zhang, Z., Dai, J.: Ghost in the Minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. arXiv preprint arXiv:2305.17144 (2023) 3