FlightGPT: Towards Generalizable and Interpretable UAV Vision-and-Language Navigation with Vision-Language Models

Hengxing Cai[1,2*], **Jinhan Dong**[2,3*], **Jingjun Tan**[1], **Jingcheng Deng**[4], **Sihang Li**[2],
Zhifeng Gao[2], **Haidong Wang**[1], **Zicheng Su**[5], **Agachai Sumalee**[6] **and Renxin Zhong**[1†]

[1]**School of Intelligent Systems Engineering, Sun Yat-Sen University** [2]**DP Technology**
[3]**Beijing University Of Posts and Telecommunications**
[4]**Institute of Computing Technology, Chinese Academy of Sciences** [5]**Tongji University**
[6]**School of Integrated Innovation, Chulalongkorn University**

Abstract

Unmanned Aerial Vehicle (UAV) Vision-and-Language Navigation (VLN) is vital for applications such as disaster response, logistics delivery, and urban inspection. However, existing methods often struggle with insufficient multimodal fusion, weak generalization, and poor interpretability. To address these challenges, we propose **FlightGPT**, a novel UAV VLN framework built upon Vision-Language Models (VLMs) with powerful multimodal perception capabilities. We design a two-stage training pipeline: first, Supervised Fine-Tuning (SFT) using high-quality demonstrations to improve initialization and structured reasoning; then, Group Relative Policy Optimization (GRPO) algorithm, guided by a composite reward that considers goal accuracy, reasoning quality, and format compliance, to enhance generalization and adaptability. Furthermore, FlightGPT introduces a Chain-of-Thought (CoT)-based reasoning mechanism to improve decision interpretability. Extensive experiments on the city-scale dataset CityNav demonstrate that FlightGPT achieves state-of-the-art performance across all scenarios, with a 9.22% higher success rate than the strongest baseline in unseen environments. Our implementation is publicly available[1].

1 Introduction

With the rapid advancement of Unmanned Aerial Vehicles (UAV) technology, vision-and-language navigation (VLN) has emerged as a critical capability for UAV applications (Fan et al., 2022; Li et al., 2025; Sautenkov et al., 2025; Wu et al., 2024). Specifically, UAV VLN involves the ability to comprehend and integrate natural language instructions with visual observations, enabling UAVs to plan and execute flight missions in complex and dynamic real-world environments (Wang et al., 2024c). This capability has demonstrated significant value across a variety of practical scenarios (Wang et al., 2024b). For example, during disaster relief operations, UAVs can rapidly identify disaster-affected areas and plan safe routes based on rescue instructions, thereby improving the effectiveness of search and rescue missions (Estrada and Ndoma, 2019).

Despite numerous methods being developed for UAV VLN task — such as sequence-to-sequence (Seq2Seq) (Fried et al., 2018), Cross-Modal Attention (CMA) (Liu et al., 2023), and Map-based Goal Predictors (MGP) (Lee et al., 2024) — several critical challenges remain in practical applications.

Insufficient multimodal information fusion. Existing methods often perform simple concatenation or shallow fusion of image and text inputs, lacking effective integration of deep semantic understanding and visual perception. Therefore, navigation strategies are prone to misinterpretation of complex instructions and perception errors, leading to suboptimal navigation performance.

Weak generalization and poor dynamic adaptability. Existing models typically rely heavily on the training environment and lack generalization capabilities in Out-of-Distribution (OOD) environments. When encountering unfamiliar environments or dynamic obstacles, their navigation performance degrades significantly, making reliable execution challenging.

Limited Interpretability of Navigation Decisions. Most current approaches directly output navigation decisions without providing clear intermediate reasoning steps. The decision-making logic is opaque to users, making it difficult to diagnose errors or refine navigation strategies, which limits the system's reliability and maintainability.

To address these challenges, we propose **FlightGPT**, a novel UAV VLN framework, as illustrated in Fig. 1. The system is built upon Vision-

---

*Equal Contribution

†Corresponding author

[1] https://github.com/Pendulumclock/FlightGPT

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6671–6688
November 4-9, 2025 ©2025 Association for Computational Linguistics

Language Models (VLMs) to support multimodal understanding, strong generalization and adaptability, and interpretable reasoning. The design of FlightGPT focuses on three techniques:

VLM-based multimodal integration. Utilizing the capacity of VLMs, visual and textual inputs are effectively integrated to enhance multimodal perception and understanding.

Two-stage training pipeline. A two-stage training pipeline is designed, starting with supervised fine-tuning (SFT) on high-quality demonstrations to warm up the policy, followed by reinforcement learning (RL) with a task-specific reward designed for UAV VLN to improve model generalization.

Chain-of-Thought based reasoning module. A structured reasoning mechanism is introduced using explicit `<think>` / `<answer>` tags, forming a Chain-of-Thought (CoT) reasoning process. This design enables the model to "think before acting" and improves reasoning quality.

The main contributions of this work are summarized as follows:

1. We leverage an end-to-end VLM that effectively integrates visual and textual inputs for enhanced multimodal comprehension.
2. A two-stage training pipeline is developed, where SFT helps accelerate convergence and stabilize training, followed by RL to enhance the model's generalization and adaptability.
3. A CoT reasoning mechanism is introduced to improve the model's reasoning quality, resulting in reasoning processes that are more complete, coherent, and fluent.
4. We evaluate FlightGPT on CityNav, a large-scale benchmark based on real-world urban environments. The model achieves state-of-the-art performance, and demonstrates strong generalization.

2 Related Work and Motivation

2.1 Evolution of UAV Vision-and-Language Navigation

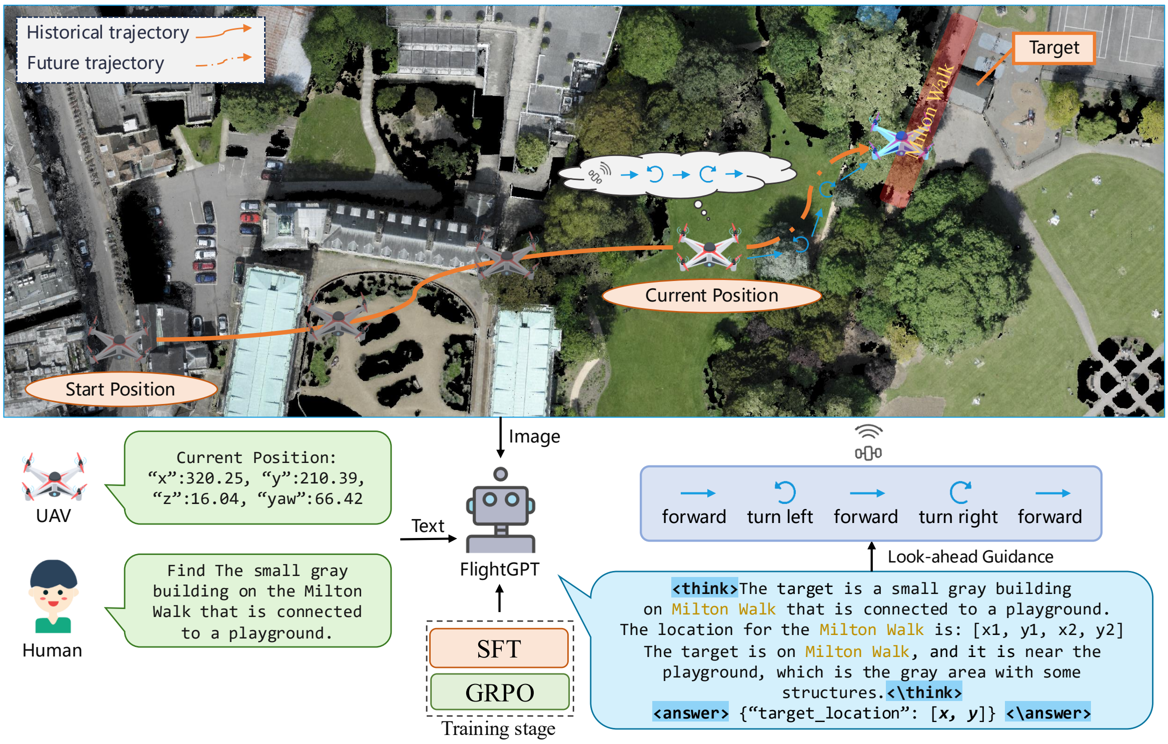
UAV VLN plays a key role in enabling intelligent flight in complex environments, and its research has undergone continuous evolution. Early UAV VLN approaches adopted Seq2Seq models that encoded language instructions into fixed representations for action generation (Fried et al., 2018). CMA mechanisms were later proposed to enhance alignment between navigation instructions and visual observations (Liu et al., 2023), while the Self-Monitoring model incorporated auxiliary progress estimation to support self-correction during navigation (Ma et al., 2019). With the rise of Transformer architectures, pretrained models such as VLN-BERT (Hong et al., 2021) were introduced, employing a multimodal BERT structure to integrate language and visual trajectories. Alongside method development, UAV VLN benchmarks have also evolved. AerialVLN (Liu et al., 2023) introduced a high-fidelity 3D simulation environment for language-guided flight, while CityNav (Lee et al., 2024) provides a city-scale dataset with GPS, imagery, and natural language, increasing task diversity and evaluation complexity. These developments have promoted the intelligent evolution of UAV VLN technologies and the standardization of benchmark datasets.

2.2 Vision-Language Models for Multimodal Perception in Navigation

VLMs, pretrained on large-scale image-text corpora, have demonstrated strong capabilities in unifying visual and linguistic modalities, making them increasingly relevant to navigation tasks that demand rich semantic perception. Early models such as UNITER (Chen et al., 2020) aligned image and text features in a joint embedding space, while CLIP (Radford et al., 2021) introduced contrastive learning for open-vocabulary visual recognition, greatly improving the generalization of multimodal representations. Recent VLMs like GPT-4V (OpenAI et al., 2024b), Gemini 1.5 (Team et al., 2024), and Qwen2-VL (Wang et al., 2024a) further expand this capability, enabling unified interfaces for vision-language reasoning and decision support. In navigation contexts, researchers have preliminarily shown that VLMs can directly process multimodal inputs to generate navigation trajectories or structured subtasks (Wang et al., 2024c). This ability to unify visual perception with language understanding positions VLMs as a promising foundation for bridging high-level task interpretation and low-level action control in navigation.

2.3 Reinforcement Learning for Enhancing Generalization in Navigation

RL has emerged as an effective mechanism for enhancing both the reasoning capabilities and generalization performance of large language mod-



Current Position: “x”:320.25, “y”:210.39, “z”:16.04, “yaw”:66.42

Find The small gray building on the Milton Walk that is connected to a playground.

<think>The target is a small gray building on Milton Walk that is connected to a playground. The location for the Milton Walk is: [x1, y1, x2, y2] The target is on Milton Walk, and it is near the playground, which is the gray area with some structures.<\think> <answer> {“target_location”: [x, y]} <\answer>

Figure 1: Workflow of FlightGPT for UAV VLN. FlightGPT takes multimodal input comprising a semantic map image and a natural language instruction, performs Chain-of-Thought reasoning to infer the target location, which is used for subsequent executable actions.

els (LLMs) and embodied agents. DeepSeek-R1 (DeepSeek-AI et al., 2025) applies large-scale RL to optimize chain-of-thought reasoning in language models, yielding strong performance in complex tasks such as mathematical problem solving and code generation. Beyond static reasoning, RL has also been leveraged to improve model adaptability in interactive settings. GROOT (Zhu et al., 2023) trains general-purpose agents in 3D environments through end-to-end RL, demonstrating the ability to generalize across diverse manipulation tasks via object-centric representations. These studies highlight the dual role of RL: not only reinforcing structured reasoning in LLMs, but also enhancing their robustness and transferability across dynamic and multi-task environments. Such capabilities are particularly valuable for UAV VLN, where agents must interpret diverse language inputs and adapt to complex, ever-changing visual contexts.

2.4 Limitations of Existing Work and Motivation for FlightGPT

While recent advances in VLMs have improved multimodal perception and RL has shown strong potential in enhancing policy generalization, their application in UAV VLN remains limited due to challenges in action reliability and training stability. To address these limitations, we propose FlightGPT, a unified framework that combines the perceptual capabilities of VLMs with the adaptive learning strengths of RL to provide a more generalizable and effective solution for UAV VLN.

3 Method

3.1 Problem Formulation

We focus on the task of UAV VLN, which requires the UAV to reach a designated target in a three-dimensional environment. The navigation process is guided by both a natural language description of the target and the UAV’s visual perception of its surroundings. Specifically, each task can be formalized as a triplet (I, D, E), where:

- I denotes the initial state of the agent, including its position and heading angle;
- D represents a natural language description of the target, typically including details about the target and its surrounding landmarks;
- E refers to a three-dimensional environment with realistic spatial layouts and rich geographic semantics, where the UAV can access various per-

ceptual inputs such as key landmarks, as well as RGB/depth maps from its first-person viewpoint.

The agent accomplishes the navigation task by executing a sequence of discrete actions, including **forwad**, **turning left**, **turning right**, **ascend**, **descend**, **stop**. When the agent determines that it has arrived near the target, it can choose the **stop** action. The navigation is considered successful if the final position of the UAV is within a predefined distance threshold (e.g., 20 meters) from the target.

3.2 FlightGPT

3.2.1 System Overview

Fig. 1 illustrates the inference process of FlightGPT when executing the UAV VLN task, which consists of the following steps:

1. Input Acquisition. The system collects inputs from the environment, including a semantic map (annotated with the UAV's current position, heading angle, first-person field of view, and known landmark information) and a textual description (containing the UAV's current position and a natural language description of the target).

2. Reasoning and Target Prediction. Following the paradigm adopted by several existing methods (e.g., Seq2Seq, CMA, MGP), we adopt a sequential workflow that first predicts the target location and then plans the navigation actions. FlightGPT generates a structured reasoning process and outputs a prediction of the target location.

3. Action Planning. Following the approach proposed in AerialVLN (Liu et al., 2023), we incorporate a look-ahead mechanism into our system, enabling the simulation of future trajectories for generating executable actions.

4. Environment Interaction. The UAV executes the planned actions in the environment and updates its state.

This process is iteratively repeated until the UAV either executes a **stop** action or reaches a predefined maximum number of iterations.

Inspired by the training process of DeepSeek-R1 (DeepSeek-AI et al., 2025), we design a two-stage training pipeline to equip *FlightGPT* with the aforementioned capabilities, as illustrated in Fig. 2.

- **Stage 1: SFT.** We use a strong model to generate training data that includes CoT-style reasoning processes. After selecting high-quality samples, we perform SFT to train the initial model. This stage is designed to provide the model with a solid initialization and to endow it with the ability to perform structured reasoning.
- **Stage 2: RL.** Building upon the SFT stage, we introduce a reinforcement learning phase based on the Group Relative Policy Optimization (GRPO) algorithm, guided by carefully designed rewards. Specifically, we define three types of rewards: Goal Accuracy Reward, Intermediate Reasoning Reward, and Format Reward. This stage aims to enhance the model's generalization ability and robustness in complex and dynamic environments.

3.2.2 SFT for Warm-up

While VLMs exhibit strong multimodal understanding, they still struggle with perception and decision-making tasks in complex and dynamic environments, such as UAV VLN. Meanwhile, RL often suffers from unstable convergence when trained from scratch. To address these challenges, we introduce a SFT stage that leverages high-quality demonstrations to warm up the model, providing a solid initialization for subsequent RL optimization.

Input, Prompt, and Output Design To enable structured output generation and strong reasoning capabilities, we design the input, prompt, and output format during the SFT stage as follows:

Input The input is composed of two parts: (1) Semantic Map, which is a map annotated with the UAV's current position and heading angle, the first-person view region projected on the map, and the locations of known landmarks; and (2) Textual Information, which describes the UAV's current state information, including its position and heading angle, along with a natural language description of the target.

Prompt To enable structured reasoning and enhance interpretability, we design a prompt template that explicitly induces a CoT style reasoning process. The prompt includes a detailed system message outlining the UAV's role and mission objective, along with structured descriptions of both the semantic map and the textual target instruction. It guides the model to reason step-by-step within dedicated <think> tags—covering semantic understanding of the target, landmark identification, and spatial inference, and to produce the final location prediction within <answer> tags. This CoT-style prompting not only improves reasoning completeness but also provides an interpretable output format that facilitates model debugging and

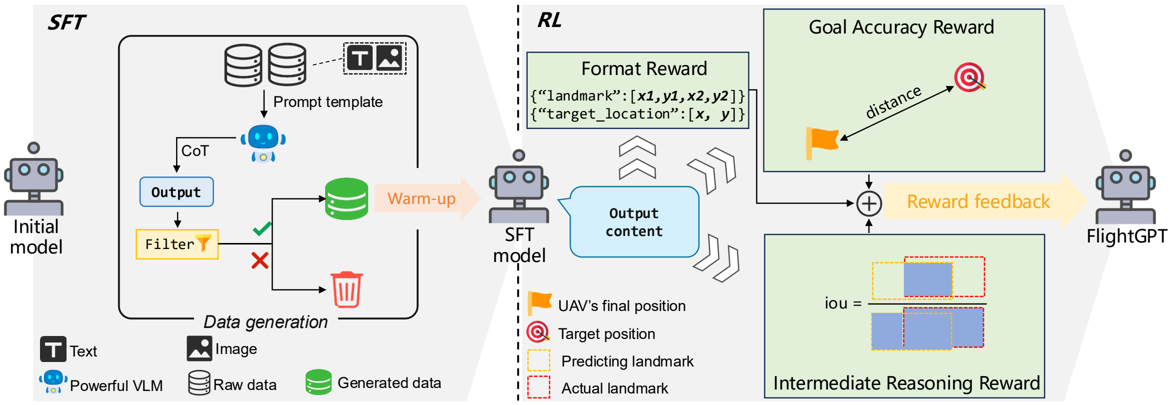


Figure 2: The two-stage training pipeline of FlightGPT. The pipeline consists of a supervised fine-tuning (SFT) stage using CoT-annotated data generated by a powerful VLM, followed by reinforcement learning (RL) with composite rewards, including goal accuracy, intermediate reasoning, and format compliance.

performance analysis. The full prompt template is provided in Appendix A.

Output The output consists of two components:

- `<think> ... </think>`: The model's intermediate reasoning process, which may include understanding the target, recognizing landmarks, and inferring spatial relationships.

- `<answer> ... </answer>`: The final predicted target location, which are used for generating subsequent executable actions.

Data Generation Due to the lack of reasoning datasets tailored for UAV VLN tasks, we adopt the Qwen2.5-VL-32B model to automatically generate the training data required for the SFT stage. Without any additional fine-tuning, we compared several open-source and closed-source VLMs, and Qwen2.5-VL-32B demonstrated the best performance under the same settings. Therefore, we select it as our data generator. Specifically, we provide Qwen2.5-VL-32B with the input and prompt template described in Section 3.2.2, guiding it to output both the reasoning process and the final prediction. To ensure the quality of the training data, we introduce the following filtering and augmentation mechanisms: (1) discard samples with abnormal output formats; (2) discard samples where the predicted location is more than 20 meters away from the ground truth; and (3) for retained samples, replace the target location predicted by Qwen2.5-VL-32B with the ground truth.

Training Strategy The training objective is next-token prediction, where the model predicts the next token based on the given input and previously generated context, proceeding token-by-token until the entire output sequence is completed.

3.2.3 RL for Generalization

Although the model acquires preliminary abilities in visual-language understanding and reasoning through the SFT stage, it still lacks the adaptability required for complex and dynamic environments—particularly in terms of generalization to unseen scenarios. To address this, we adopt the GRPO algorithm to perform policy optimization using pre-collected simulated data, based on the multimodal input and prompt templates defined in Section 3.2.2.

To jointly improve final navigation accuracy, reasoning quality, and output format consistency, we design a composite reward system consisting of the following three components:

Goal Accuracy Reward. The accuracy of the predicted target location is a key indicator of the system's effectiveness. Let the UAV's predicted position be $\hat{\mathbf{p}} = (\hat{p}_x, \hat{p}_y)$ and the ground truth be $\mathbf{p}^* = (p_x^*, p_y^*)$. We define the reward based on their distance:

$$R_{\text{goal}} = \begin{cases} 1, & \text{if } d(\hat{\mathbf{p}}, \mathbf{p}^*) \leq d_{\text{success}} \\ \exp\left(-\dfrac{d(\hat{\mathbf{p}}, \mathbf{p}^*) - d_{\text{success}}}{\tau}\right), & \text{if } d_{\text{success}} < d(\hat{\mathbf{p}}, \mathbf{p}^*) \leq d_{\text{cutoff}} \\ 0, & \text{otherwise} \end{cases}$$

where:

- The Euclidean distance $d(\hat{\mathbf{p}}, \mathbf{p}^*)$ is defined as:

$$d(\hat{\mathbf{p}}, \mathbf{p}^*) = \sqrt{(\hat{p}_x - p_x^*)^2 + (\hat{p}_y - p_y^*)^2}.$$

- $d_{\text{success}} = 20$ meters: threshold for task success;

- $d_{\text{cutoff}} = 80$ meters: upper limit beyond which no reward is given;
- $\tau = 100$: decay temperature controlling the sharpness of the exponential drop-off.

This reward encourages the model to generate target locations that are closer to the ground truth, thereby improving success rate.

Intermediate Reasoning Reward. Providing guidance for intermediate reasoning steps is critical for enhancing multi-step navigation performance. In our task, we leverage landmarks as key intermediate signals to encourage effective reasoning during the `<think>` stage. Specifically, we introduce a reward based on the Intersection over Union (IoU) between the predicted landmark bounding box $\hat{B}$ and the ground-truth bounding box B. The reward is defined as:

$$R_{\text{IoU}} = \frac{\text{Area}(B \cap \hat{B})}{\text{Area}(B \cup \hat{B})}$$

If the model fails to output a valid bounding box, we set $R_{\text{IoU}} = 0$. This mechanism incentivizes spatial reasoning before location prediction and contributes to more stable and interpretable intermediate representations.

Format Reward. To ensure the model generates structured outputs, we introduce a format compliance reward. This reward encourages the model to produce outputs that follow a predefined format, with both the reasoning and action sections clearly presented and containing the required information.

- If the output includes both `<think>` and `<answer>` tags properly enclosing the reasoning and answer segments, a reward of $+0.5$ is given;
- If a `"landmark_bbox"` field in the format `[x1, y1, x2, y2]` is successfully extracted within the `<think>` tag, an additional $+0.25$ is granted;
- If a `"target_location"` field in the format `[x, y]` is successfully extracted within the `<answer>` tag, another $+0.25$ is granted.

This reward helps stabilize the model's output structure, facilitating controllability and enabling downstream execution or interpretation.

Overall Reward. The total reward used for policy optimization is the sum of the three components described above:

$$R_{\text{total}} = R_{\text{goal}} + R_{\text{IoU}} + R_{\text{format}}$$

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

In this study, we utilize the CityNav (Lee et al., 2024) dataset, a high-quality benchmark specifically designed for city-scale UAV VLN tasks. CityNav comprises 32,637 human demonstration trajectories across 5,850 target objects, constructed on top of 3D urban scans from the SensatUrban dataset. It covers two real-world cities, Birmingham and Cambridge, with a total area of approximately 4.65 km², providing rich geographic semantics and diverse navigation scenarios. To better reflect UAV flight conditions, CityNav introduces altitude variation in its task design. At the beginning of each episode, the UAV is randomly initialized at heights below 150 meters, and the inputs include first-person-view images captured from different altitudes together with the corresponding height values. The dataset is publicly available under the MIT License, enabling free use for research purposes.

4.1.2 Evaluation Metrics

Following the standard evaluation protocol established by CityNav, four metrics are used to evaluate performance:

- **Navigation Error (NE)**: The Euclidean distance between the agent's final position and the ground-truth location. Lower NE indicates better localization accuracy.
- **Success Rate (SR)**: The percentage of episodes in which the agent stops within 20 meters of the target location.
- **Oracle Success Rate (OSR)**: The proportion of episodes where the agent, at any point during navigation, gets within 20 meters of the target, regardless of whether it stops.
- **Success weighted by Path Length (SPL)**: A metric that adjusts SR by penalizing unnecessarily long paths, encouraging efficient navigation.

These metrics jointly reflect the agent's goal-reaching accuracy, path efficiency, and overall navigation robustness.

4.1.3 Baseline Models

We conduct evaluations of FlightGPT against a diverse set of representative baselines, including Random, Seq2Seq, CMA, MGP, GPT-4o, Qwen2.5-VL

Table 1: Hyperparameters for SFT and RL stages.

Stage	Batch Size	LR	Epochs
SFT	16	2e-5	2
RL	1	1e-5	1

(7B / 32B), and LLaMA-3.2-11B-Vision. Brief introductions for all baselines are provided in Appendix B.

4.1.4 Model and Training Configuration

We adopt Qwen2.5-VL-7B as the base model and optimize it using a two-stage pipeline. For SFT, 1,872 samples were collected and filtered from Qwen2.5-VL-32B outputs. For RL, 4,758 samples were selected from the training set, covering diverse cities, street scenes, and target types. The SFT stage is implemented using **LLaMA-Factory** (Zheng et al., 2024), while the RL stage is built upon the **VLM-R1** framework (Shen et al., 2025). Key hyperparameters for both stages are summarized in Table 1.

4.2 Experimental Results

4.2.1 Model Performance and Generalization Analysis

Table 2 summarizes the performance of various models across evaluation scenarios in the CityNav dataset. Experimental results reveal that Qwen2.5-VL-7B achieves reasonable performance in UAV VLN tasks, while its larger variant, Qwen2.5-VL-32B, further improves and surpasses the strongest traditional baseline, MGP, across multiple metrics. These observations underscore that base VLMs already possess strong visual-language perception and multimodal fusion capabilities, even when used out-of-the-box without task-specific tuning.

On top of this foundation, FlightGPT further improves performance across the board. In the val-seen setting, it achieves the highest success rate **17.57%**, the lowest navigation error **66.1**, and the most efficient path SPL **15.78**. In more challenging test-unseen setting, it shows remarkable generalization ability, improving the success rate by **9.22%** and nearly doubling the SPL compared to Qwen2.5-VL-32B, the strongest baseline model.

It is worth noting that FlightGPT, built on the relatively lightweight Qwen2.5-VL-7B model, surpasses the larger-scale Qwen2.5-VL-32B after the application of a two-stage training pipeline. This result highlights that, rather than merely scaling up model size, incorporating appropriate modeling approaches (e.g., a CoT reasoning module) and adopting efficient training strategies (e.g., SFT+RL) are more crucial for enhancing model generalization and real-world performance.

4.2.2 Training Strategy Analysis

To systematically evaluate the contributions of SFT and RL in the FlightGPT framework, we conduct ablation experiments under the following three training configurations: (1) **SFT-only**: Trained with supervised fine-tuning only, without RL; (2) **RL-only**: Trained directly with reinforcement learning, without SFT initialization; (3) **SFT+RL**: Initialized with SFT and further optimized with RL.

- **SFT-only**: This configuration achieves decent performance in the val-seen environment, benefiting from the reasoning mechanism and SFT on high-quality data. However, without RL for policy optimization and exploration, it shows limited generalization to OOD environments. On the test-unseen set, its performance is clearly inferior to models trained with RL.

- **RL-only**: This configuration eventually achieves reasonably good performance without any prior initialization. However, as shown in Fig. 3, the model suffers from low success rates at the beginning of training due to the absence of a good starting policy. Its convergence is slower than SFT+RL: while SFT+RL nearly converges at around 500 steps, the RL-only model only begins to stabilize after 600 steps, and its reward remains consistently lower throughout training. In addition, its final performance remains slightly lower than that of SFT+RL.

- **SFT+RL**: The SFT stage provides a strong initialization of the policy, resulting in a more stable and faster convergence during training. Subsequently, the RL stage further improves the model's generalization and adaptability to OOD environments. This configuration not only outperforms both SFT-only and RL-only baselines across all evaluation metrics, but also achieves a more stable and efficient training process, demonstrating the synergistic advantage of the two-stage training pipeline.

4.2.3 Reward Component Ablation

To further understand the contribution of each reward design, we conduct ablation experiments by

Table 2: Comparison of Model Performance Across Evaluation Scenarios

Method	**Validation Seen**				**Validation Unseen**				**Test Unseen**			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random	222.30	0.00	1.15	0.00	223.00	0.00	0.90	0.00	208.80	0.00	1.44	0.00
Seq2Seq	148.40	4.52	10.61	4.47	201.40	1.04	8.03	1.02	174.50	1.73	8.57	1.69
CMA	151.70	3.74	10.77	3.70	205.20	1.08	7.89	1.06	179.10	1.61	10.07	1.57
MGP	59.70	8.69	**35.51**	8.28	75.10	5.84	22.19	5.56	93.80	6.38	26.04	6.08
Qwen2.5-VL-7B	116.10	4.72	12.89	4.15	123.20	5.52	13.98	4.92	124.60	4.59	12.75	3.99
Qwen2.5-VL-32B	84.70	12.65	24.14	11.30	91.90	10.12	20.52	9.00	83.28	11.98	23.48	10.76
LLaMA-3.2-11B-Vision	198.90	1.16	5.16	1.06	215.10	0.50	4.35	0.46	191.10	1.26	4.59	1.15
GPT-4o	155.80	2.42	9.62	2.17	170.40	2.17	7.77	1.98	144.40	3.90	11.79	3.42
SFT-only	97.60	10.29	18.45	9.46	101.70	10.51	18.54	9.70	117.40	11.20	21.24	10.78
RL-only	74.90	13.27	27.13	12.59	71.40	12.87	25.82	12.27	76.50	19.80	32.26	18.91
SFT+RL (FlightGPT)	**66.10**	**17.57**	30.26	**15.78**	**68.10**	**14.69**	**29.33**	**13.24**	**76.20**	**21.20**	**35.38**	**19.24**

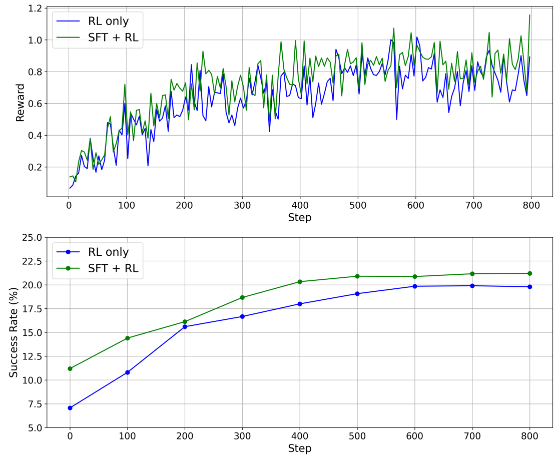


Figure 3: Reward (train) and success rate (test) over training steps.

individually removing components from the composite reward. The results on the Test Unseen set are summarized in Table 3. We observe that:

- Removing the **Goal Accuracy Reward** causes the most significant drop in navigation performance, confirming its pivotal role in guiding the model toward correct target localization.
- The **Intermediate Reasoning Reward** helps the model better utilize landmark-based spatial reasoning during the `<think>` stage, improving intermediate predictions.
- The **Format Reward** strengthens the consistency and regularity of the model outputs, thereby improving the overall reliability of the generated plans.

Overall, these components are complementary and together lead to a substantial performance boost, indicating strong synergy among them.

4.2.4 Reasoning Quality Analysis

To compare the reasoning quality between the RL-only and SFT+RL configurations, we randomly selected several cases from the dataset for qualitative analysis. The reasoning process generated by the RL-only model is generally disorganized: the `<think>` section tends to be short, lacks clear logical structure, and contains fragmented reasoning chains, making it difficult to follow. In contrast, SFT+RL produces significantly more coherent and well-structured reasoning, with complete chains covering landmark identification, spatial relation reasoning, and target location prediction. Several representative examples are provided in Appendix C.

To further quantify these observations, we designed three reasoning quality metrics and used GPT-4o to automatically score a random sample of 5,000 outputs (the detailed prompt is provided in Appendix D). The three reasoning quality metrics are measured by: (1) **Completeness**: Whether the reasoning covers all necessary steps without missing key details; (2) **Coherence**: Whether the reasoning is logically consistent and well connected throughout; (3) **Fluency**: Whether the language is fluent and grammatically correct.

To reduce evaluation variance, each sample was scored 3 times, and the average score was reported as the final result. The evaluation results, summarized in Table 4, show that SFT+RL outperforms the RL-only model across all three reasoning quality metrics, demonstrating the critical role of the SFT stage in improving reasoning quality. In particular, the SFT+RL configuration achieves a 0.44 improvement in completeness, indicating that structured reasoning training effectively guides the model to produce more comprehensive and systematic reasoning processes. Additionally, improvements of 0.26 and 0.08 are observed in coherence and fluency, respectively, further enhancing the clarity and readability of the reasoning outputs.

Table 3: Ablation results of reward components on the Test Unseen set.

Method	**NE↓**	**SR↑**	**OSR↑**	**SPL↑**
FlightGPT	76.20	21.20	35.38	19.24
w/o Goal Accuracy Reward	102.02	9.07	14.37	8.30
w/o Intermediate Reasoning Reward	77.89	18.34	31.75	16.81
w/o Format Reward	77.56	19.20	27.60	17.50

Table 4: Reasoning Quality Evaluation Results

Strategy	**Completeness**	**Coherence**	**Fluency**
RL-only	3.67	4.03	4.78
SFT+RL	4.11	4.29	4.86

4.2.5 Inference Efficiency Analysis

Efficient inference is crucial for real-world deployment of UAV VLN, particularly in resource-constrained edge environments. To assess the practicality of FlightGPT, we evaluated its memory usage and inference latency during the deployment phase and compared them with several representative VLMs. The results are summarized in Appendix E. On a single RTX 4090 GPU, FlightGPT exhibits an average latency of 9.37 seconds per step and memory usage of 21.71 GB. Compared with other VLMs, FlightGPT achieves relatively higher inference efficiency, which suggests strong potential for edge deployment.

5 Conclusion

In this paper, we propose FlightGPT, a system for UAV VLN, aiming to improve navigation performance in complex environments, enhance cross-environment generalization, and increase the interpretability of decision-making processes. We leverage the multimodal understanding capabilities of VLMs and construct a two-stage training pipeline that combines SFT with RL, where the RL stage is guided by a composite reward design. In addition, we introduce a CoT reasoning mechanism to improve the transparency and controllability of the system. We conduct comprehensive evaluations on the real-world, city-scale CityNav dataset. Experimental results show that FlightGPT achieves significant improvements over existing baseline models in in-distribution environments, and exhibits strong generalization capabilities in more challenging OOD scenarios. We will release the code and data to facilitate further research.

6 Limitations

Despite the strong performance of FlightGPT in city-scale UAV VLN, several noteworthy limitations remain in terms of real-world applicability and system capabilities:

Significant Gap Between Simulation and Reality. This study primarily relies on high-fidelity simulators such as CityNav for training and evaluation. While these platforms offer structured and diverse urban scenarios that facilitate learning of task semantics and spatial layouts, they fall short of capturing the complexity and uncertainty of real-world urban airspaces. Factors such as GPS drift, weather disturbances, dynamic obstacles, and unexpected events frequently arise in actual UAV operations and can significantly impact perception and decision-making. As a result, the system's performance, stability, and robustness in real-world settings remain unverified and call for further field testing and validation.

Substantial Gap Compared to Human Navigation Abilities. Although FlightGPT demonstrates leading performance on the CityNav dataset and exhibits basic language understanding and path planning capabilities, its navigation intelligence still lags behind human operators. In particular, the model struggles with complex scenarios involving ambiguous expressions, implicit goals, or multi-turn instructions, often lacking commonsense reasoning and strategic flexibility. This exposes limitations in multi-modal semantic integration, spatial reasoning, and decision consistency, making it difficult for the system to handle dynamic and high-complexity navigation tasks.

Lack of Systematic Evaluation of Deployment Feasibility. The current research primarily emphasizes performance, with insufficient attention paid to the practical requirements for real-world deployment. Key factors such as inference latency, memory usage, and computational resource demands directly influence the system's ability to operate in real time on resource-constrained edge devices, yet these metrics have not been systematically quantified. Furthermore, issues such as communication reliability and failure recovery mechanisms—critical for engineering-level implementation—remain underexplored, limiting the transition of FlightGPT from a research prototype to a deployable solution.

Challenges in Handling Complex Instructions. While FlightGPT is able to follow basic navigation commands, it lacks explicit mechanisms for staged or hierarchical planning, which limits its ability to process instructions with semantic progression or conditional triggers. This shortcoming makes it difficult for the system to reliably execute multi-step goals and capture dependencies across intermediate sub-tasks, reducing its effectiveness in scenarios that require more structured and adaptive reasoning. Future research could address this issue by incorporating modular planning frameworks or hierarchical understanding mechanisms, enabling the model to better represent and carry out complex, structured instructions.

7 Broader Impact and Ethics

Dual-use risk. UAV-based navigation systems, while beneficial for disaster relief or infrastructure inspection, may also be misused for surveillance, tracking, or other purposes that infringe on privacy or civil liberties. To mitigate such risks, real-world deployment should be accompanied by appropriate regulatory oversight, strict usage boundaries, and human-in-the-loop supervision mechanisms.

Risk of unsafe deployment. Although the system shows strong performance in simulated city-scale environments, deploying it in real-world scenarios poses safety risks due to unmodeled factors such as GPS drift, occlusions, dynamic obstacles, or weather conditions. Without rigorous field testing and fail-safe mechanisms, these issues may lead to unintended navigation failures or even physical harm to people or property.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, and Anton Van Den Hengel. 2017. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. *Preprint*, arXiv:1909.11740.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mario Arturo Ruiz Estrada and Abrahim Ndoma. 2019. The uses of unmanned aerial vehicles–uav's-(or drones) in social logistic: Natural disasters response and humanitarian relief aid. *Procedia Computer Science*, 149:375–383.

Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Eric Wang. 2022. Aerial vision-and-dialog navigation. *arXiv preprint arXiv:2205.12219*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vlnbert: A recurrent vision-and-language bert for navigation. In *Computer Vision and Pattern Recognition*.

Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr deghosting from frequency view. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. 2024. Citynav: Language-goal aerial navigation dataset with geographic information. *Preprint*, arXiv:2406.14240.

Ye Li, Li Yang, Meifang Yang, Fei Yan, Tonghua Liu, Chensi Guo, and Rufeng Chen. 2025. Navblip: a visual-language model for enhancing unmanned aerial vehicles navigation and object detection. *Frontiers in Neurorobotics*, 18:1513354.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yaning Zhang, and Qi Wu. 2023. Aerialvln: Vision-and-language navigation for uavs. *Preprint*, arXiv:2308.06735.

Chih Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan Alregib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai

Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Oleg Sautenkov, Yasheerah Yaqoot, Artem Lykov, Muhammad Ahsan Mustafa, Grik Tadevosyan, Aibek Akhmetkazy, Miguel Altamirano Cabrera, Mikhail Martynov, Sausar Karaf, and Dzmitry Tsetserukou. 2025. Uav-vla: Vision-language-action system for large scale aerial mission generation. *arXiv preprint arXiv:2501.05014*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina

Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan

Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Tian Wang, Pai Zheng, Shufei Li, and Lihui Wang. 2024b. Multimodal human–robot interaction for human-centric smart manufacturing: a survey. *Advanced Intelligent Systems*, 6(3):2300359.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2024c. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*.

Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. 2024. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. 2023. Learning generalizable manipulation policies with object-centric 3d representations. *Preprint*, arXiv:2310.14386.

A Prompt Template for FlightGPT

Prompt

System Message:
You are an intelligent autonomous aerial vehicle (UAV) capable of real-world navigation and visual target localization.

Mission Objective:
Your mission is to locate a specific target described in natural language instructions.

Details of the Target:
`{target description}`

Environmental Perception:
- The UAV's current position is indicated by the starting point of an arrow in the image, with its heading angle represented by the arrow's direction.
- The yellow box outlines the UAV's current camera field of view on the map, centered at pixel coordinates: `cur_pose = {UAV current position}`.
- Landmark regions are highlighted with red masks.

Operational Guidance:
- The target is usually located near a red-masked landmark.
- Use both the target description and the visual input to identify the most relevant red-masked landmark region.
- Infer the relative position of the target with respect to that landmark.

Output Format Specification:
- Present your reasoning process within `<think>` and `</think>` tags.
- Provide your final answer within `<answer>` and `</answer>` tags in the following format: `{"target_location": [x, y]}`
Your reasoning may include:
- A semantic interpretation of the target description.
- Identification of the correct landmark region.
- The bounding box of that region in the following format:
`{"landmark_bbox": [x1, y1, x2, y2]}`

B Baseline Model Descriptions

We briefly introduce the baseline models evaluated on the CityNav.

- **Random**: A simple random action policy serving as a lower bound reference.
- **Seq2Seq** (Anderson et al., 2017): A classic end-to-end model that encodes the instruction and visual input into a latent representation to directly generate action sequences.
- **CMA** (Hu et al.): A cross-modal attention model that integrates visual and linguistic inputs via cross-attention mechanisms for action selection.
- **MGP** (Lee et al., 2024): A map-based method that predicts the target location based on semantic maps and natural language instructions.
- **GPT-4o** (OpenAI et al., 2024a): A powerful multimodal model developed by OpenAI, capable of processing both text and visual inputs with reasoning abilities.
- **Qwen2.5-VL (7B / 32B)** (Bai et al., 2025): An open-source family of vision-language models with varying parameter scales, used to assess how model size influences performance.
- **LLaMA-3.2-11B-Vision** (Grattafiori et al., 2024): Meta AI's latest open-source multimodal

model, supporting both visual and language perception.

C Reasoning Examples

D Prompt for Reasoning Quality Evaluation

Prompt

System Message:
You are a language model output quality evaluator. Please evaluate the following text based on the three criteria below.

Evaluation Criteria Definitions:
1. Completeness
- Does the text fully present all the necessary reasoning steps or information needed to answer the question or solve the problem?
- Are there any missing steps, skipped logic, or gaps in reasoning?
2. Coherence
- Is the reasoning internally consistent and logically connected? Are there any contradictions or abrupt jumps?
3. Fluency
- Is the language natural and smooth? Are grammar, sentence structure, and word choice appropriate?

Scoring Guideline:
- 1: Very poor
- 2: Poor
- 3: Fair
- 4: Good
- 5: Excellent

Based on the above definitions and scoring scale, please evaluate the following text. Output one integer score (1–5) per criterion. Strictly follow the format below so it can be parsed by code.

Language Model Output:
`{language model output}`

Output Format:
- Completeness: x
- Coherence: x
- Fluency: x

E Resource Consumption During Inference

Table 5: Reasoning Examples

ID	RL-only Output (<think> Section)	SFT+RL Output (<think> Section)
1	The landmark region is the area where the red mask is present. The bounding box for the red-masked region is: `{"landmark_bbox": [3428, 23, 3801, 547]}` The probable target location is on the sidewalk next to the building within this red-masked area.	The target is a black car on the sidewalk next to a building on Chesterwood Gardens. The red masks in the image indicate the streets, specifically Chesterwood Gardens, and may help us identify the correct landmark region for locating the target. **1. Semantic Interpretation of the Instruction:** – The car is described as a black car and is on the sidewalk near a building. – The location reference “Chesterwood Gardens” should point the drone towards the correct residential area. **2. Identifying the Correct Landmark Region:** – The aerial view shows Chesterwood Gardens as a rectangular area marked by the red mask. Given it is residential with houses and likely the target is near the edge of one of these buildings on the sidewalk, I would focus on the northern edge of Chesterwood Gardens. **3. Relative Position of the Target:** – The target is described as on the sidewalk next to a building. The sidewalk along the northern edge of Chesterwood Gardens is a likely location. **4. Bounding Box for the Landmark Region:** – Based on the aerial view, the northern edge of Chesterwood Gardens is highlighted in red. The area around the edge of the buildings and the sidewalk here aligns with the target description. `{"landmark_bbox": [2693, 16, 4026, 1361]}`
2	The target, a black car parked between a silver car and a black car, is located in the parking lot behind the "One Stop" store. The red-masked landmark represents the One Stop store, which is a good reference point. The target is likely to be in the parking lot area adjacent to this landmark, as indicated by the "One Stop" text. The parking lot is clearly demarcated by the red masks. Given the position of the target relative to the One Stop landmark, it should be in the area where the parking rows are arranged around the store. Since the target is described as being between two cars, we can look for that specific arrangement of cars in the parking lot area behind One Stop. Since the target is in the parking lot area behind One Stop, the bounding box would cover the parking lot region near the red-masked "One Stop" area. `{"landmark_bbox": [1745, 0, 3882, 1467]}` The target, a black car, is likely to be found within the bounding box of the parking lot area behind "One Stop".	The target description states that it is a black car parked between a silver car and a black car in a lot behind "One Stop." Given the current position indicated by the arrow, we need to identify a red-masked landmark near this area as the target might be near a red-masked street landmark for reference. **1. Identify Possible Red-Masked Landmarks:** – The image highlights red structures with some marked as "One Stop," which appears larger and more prominent. – The area behind "One Stop" is clearly outlined by the red mask. **2. Locate the Target Relatively:** – The target is described as being behind "One Stop." – Given the current field of view and the position indicated by the arrow, the target should be within the vicinity of the "One Stop" landmark. **3. Bounding Box for Landmark:** – The "One Stop" landmark is clearly marked by the red structure. – The bounding box for the landmark can be identified roughly around the "One Stop" structure. `{"landmark_bbox": [1594, 0, 3864, 1502]}`
3	The target description indicates that the target is a parking lot located behind the building at the corner of Chesterton Road and Ferry Path. The red mask outlines the street landmark regions. The target should be in the parking lot behind the building at the intersection of Chesterton Road and Ferry Path. Let’s identify the landmark region: – The red mask around Chesterton Road and Ferry Path clearly highlights the region where the target is likely to be located. – The target parking lot is behind the building at the corner of Chesterton Road and Ferry Path. The bounding box for the landmark region is: `{"landmark_bbox": [403, 249, 1060, 959]}` Given the green arrow and the current orientation of the UAV, the target is likely to be behind the building at the corner of Chesterton Road and Ferry Path within the red mask area.	The task is to find a parking lot, which is described as "behind the building at the corner of Chesterton Road and Ferry Path" and "has no cars in this area." The goal is to locate this parking lot, relative to the red-masked landmark regions marked as "Chesterton Road" and "Ferry Path". **1. Identify the relevant landmark red masks:** – There are red masks over Chesterton Road and Ferry Path in the image. These identify the streets and regions of interest. **2. Determine the direction of movement:** – The UAV’s current orientation is indicated by the arrow at the bottom-left. It points towards the top-left of the image. **3. Locate the intersection:** – The intersection of Chesterton Road and Ferry Path is where the red masks overlap at the bottom of the image. **4. Identify the target area:** – The target parking lot is described as being behind the buildings at the corner. Given the red masks, it’s reasonable to conclude that any area behind these masks, specifically in the region near the intersection, is the location to look for. **5. Bounding box of the area:** – Since we need to look behind the intersection at the parking lot, and considering the red masks as a guide, I’ll assume an area directly behind the intersection is our target. This means the parking lot would lie inside the yellow bounding box (highlighted on the image). The bounding box region of interest behind the red-masked Chesterton Road, which should contain the parking lot, would be: `{"landmark_bbox": [370, 273, 1076, 972]}`

Table 6: Resource consumption during inference.

Model Name	Test Device	Memory Usage (GB)	Inference Latency (s/step)
FlightGPT (7B)	RTX 4090 ×1	21.71	9.37
LLaMA-3.2-11B-Vision	RTX 4090 ×1	21.65	11.11
Qwen2.5-VL-32B	A100 80GB ×1	70.12	53.42
GPT-4o	Cloud	N/A	9.73