Game Ground Bench: Probing the Limits of LVLMs in Complex Semantic Grounding Across Game Universes

Zhangyang Qi[1, 2], **Jinsong Li**[2], **Hongjian Wu**[1], **Jiaqi Wang**[2*], **Hengshuang Zhao**[1*]
[1]The University of Hong Kong,
[2]Shanghai Artificial Intelligence Laboratory
{zyqi,hszhao}@cs.hku.hk, {lijingsong,wangjiaqi}@pjlab.org.cn

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities, yet their ability to ground language in complex, interactive environments such as video games remains a critical frontier. Existing benchmarks are inadequate for this purpose: real-world datasets like RefCOCO introduce a domain gap; GUI-centric benchmarks lack the complexity of modern game interfaces; and existing game-specific benchmarks are often too simplistic or narrow, failing to assess fine-grained, generalizable grounding capabilities. To address this issue, we propose GGBench — a large-scale, cross-genre benchmark designed to probe the grounding capabilities of LVLMs in diverse gaming scenarios. GGBench features unprecedented genre diversity, encompassing 10 categories including card games, first-person shooters, and role-playing games, with a total of 1335 test images. It focuses on tasks that require connecting natural language instructions to specific in-game objects and UI elements. Experimental results show existing models perform poorly on GGBench, with weak grounding abilities, especially in complex game scenarios. Due to limited data scale, fine-tuning them for gaming scenarios is also challenging. To address this, we propose Game-R1, a novel training method centered on the Grounded Reinforcement Policy Optimization (GRPO) algorithm. GRPO maximizes limited interaction data utility and enables robust few-shot generalization across games. Extensive experiments show Game-R1 significantly outperforms existing LVLMs on GGBench, validating our approach. GGBench provides a solid and comprehensive evaluation platform for subsequent research on agents in gaming environments, which strongly promotes development in this field.

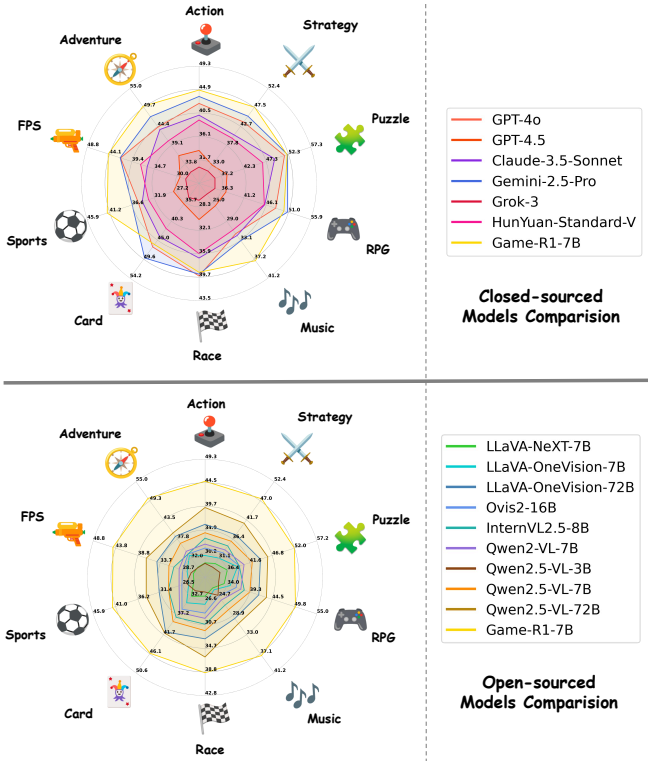

Figure 1: **GGBench Radar Chart**. This chart shows various LVLMs' performance on GGBench. Overall, closed-source models outperform open-source ones. Notably, Game-R1, after 500-sample Reinforcement Fine-tuning, surpasses closed-source models—validating the small-batch training.

1 Introduction

Large Vision-Language Models (LVLMs) have advanced rapidly, transforming digital interaction with their powerful capabilities in understanding and generating natural language. **Video games**, with their massive user bases and high interactivity, serve as an ideal testing ground for these models by providing complex, dynamic environments to deploy AI agents. For true human-computer collaboration in these virtual spaces, the **Grounding** task is critical. This requires an agent to not only grasp textual commands but also connect that abstract language to specific in-game objects or regions. Accurate grounding enables effective navigation, interaction, and task execution, forming the foundation for autonomous intelligence in games.

Current benchmarks for evaluating grounding capabilities suffer from critical limitations. **Mainstream grounding benchmarks** like RefCOCO (Kazemzadeh et al. 2014), which leverage real-world images, have driven progress in vision-language alignment but inherently create a domain gap with virtual game environments—casting doubt on their

*Corresponding Author

Figure 2: **Example of GGBench.** Here is one example from each of the ten game categories. Our images are high-definition game screenshots, with indirect questions requiring LVLMs to conduct in-depth reasoning on image content, including retrieving in-game objects and identifying UI buttons.

ability to assess model applicability in gaming contexts. **GUI-based benchmarks** (e.g., those for Screenspot (Cheng et al. 2024)) are indeed designed to evaluate LVLMs in the context of computer interface grounding. However, their interfaces are characterized by standardized, simplistic structures—failing to capture modern game interfaces' intricate complexity, making them ill-suited for assessing model performance in gaming scenarios. **Gaming-specific benchmarks** even fall short: many focus on simplistic games (e.g., 2048) (Qin et al. 2025), while those tied to a single platform (e.g., Atari) (Mnih et al. 2013) further impede generalization and fail to align with modern commercial games. Moreover, these benchmarks narrow their evaluations to outcomes like wins or losses, neglecting critical grounding in in-game objects or functional controls. Therefore, we introduce ***GGBench***: a LVLM benchmark specifically designed to evaluate grounding capabilities in complex and diverse scenarios.

GGBench is a large-scale and cross-genre gaming benchmark designed to evaluate an AI agent's grounding ability in complex visual environments. It focuses on two key tasks: locating objects from textual descriptions and selecting the correct UI element from instructions. The primary advantage is the unprecedented genre diversity, including **card games** that test strategic UI comprehension (e.g., *Hearthstone*), **sports games** involving dynamic target tracking (e.g., *Mario Tennis*), **first-person shooters (FPS)** requiring precise localization in fast-paced 3D environments (e.g., *CS-Go*), **racing games** that focus on dashboard and environmental recognition (e.g., *Asphalt*), **music rhythm games** that test sequential visual response (e.g., *Taiko no Tatsujin*), and strategy and **role-playing games (RPG)** with high-density, complex UIs (e.g., *League of Legends*). There are 10 categories in total and 1335 images. GGBench aims to offer a more comprehensive, equitable platform for assessing AI agents' true grounding capabilities in complex interactive environments, advancing related research.

Creating a benchmark as comprehensive and diverse as GGBench challenges existing LVLMs, particularly learn-

ing across distinct game environments with limited labeled data per domain. We propose ***Game-R1***, a novel training method designed to develop robust, generalizable grounding capabilities in few-shot or limited-data scenarios to tackle this data-efficient learning challenge. Central to Game-R1 is GRPO (Grounded Reinforcement Policy Optimization), an advanced policy optimization algorithm that maximizes limited interaction data utility, enabling fast convergence and effective generalization from few samples. We hypothesize GRPO's high sample efficiency will make Game-R1 outperform other LVLMs on GGBench.

- **New testing benchmark (GGBench)**: A multi-game benchmark evaluating AI's ability to map language instructions to in-game visuals in diverse scenarios, filling the gaming domain's grounding evaluation gap.
- **Efficient training method (Game-R1)**: Proposed Game-R1, centered on the GRPO algorithm. It handles cross-game learning with limited data, enabling AI to accurately understand complex game interfaces.
- **Remarkable experimental results**: Tests on GGB show that Game-R1 significantly outperforms other models, while keep the VQA and OCR capability.

2 Related work

LVLMs For Grounding Task. The visual grounding capabilities of Large Vision-Language Models (LVLMs) have evolved significantly from holistic semantic understanding to fine-grained region-level localization. Initial efforts, such as BuboGPT (Zhao et al. 2023), adopted an "off-the-shelf module" strategy, relying on separate visual models to locate entities, which limited deep cross-modal fusion. To achieve tighter vision-language integration, mainstream research has rapidly shifted towards end-to-end solutions that make grounding an inherent model capability. These solutions primarily follow two technical paradigms: the "pixel-to-sequence" (pix2seq) approach, seen in models like Kosmos-2 (Peng et al. 2024) and GroundingGPT (Li et al. 2024b), which serialize coordinates into text tokens for unified prediction; and the "pixel-to-embedding" (pix2emb) approach, pioneered by NExT-Chat (Zhang et al. 2024), which outputs location embeddings that can be decoded into bounding boxes or even pixel-level masks, as demonstrated by GLaMM (Rasheed et al. 2024). Furthermore, addressing the demands of dynamic scenes like games, cutting-edge works such as VTimeLLM (Huang et al. 2024) have successfully extended grounding from the spatial to the temporal domain across video and audio, laying a solid foundation for precise interaction in complex multimodal environments.

Gaming-Benchmark. Games have long served as a classic testbed for AI agents, particularly in reinforcement learning, where "Grand Challenges" built around single, complex environments like StarCraft II (Vinyals et al. 2019), Dota 2 (OpenAI 2019), and Minecraft (Baker et al. 2022) have successfully advanced the development of specialist agents. However, these benchmarks, with their focus on depth within a single domain, are ill-suited for evaluating the generalizable capabilities across diverse tasks—a core strength of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). Although the community has begun to adopt games for evaluating large models, existing approaches have notable limitations (Hudi et al. 2025). They are often confined to either text-only interactions, which preclude the assessment of crucial vision-language grounding capabilities (Yu et al. 2025), or they incorporate vision merely through static Visual Question Answering (VQA) tasks (Tsai et al. 2023), failing to evaluate an agent's ability to make sequential decisions in dynamic environments. There is, therefore, a pressing need for a benchmark that spans multiple universal game types to comprehensively probe the grounding capabilities of LVLMs in novel, interactive settings.

GRPO-RL for Training LVLMs. Inspired by the success of GRPO-based reinforcement learning in LLMs like DeepSeek-R1 (Guo, Zhang et al. 2025), this paradigm has been increasingly applied to Large Vision-Language Models (LVLMs). Initial explorations such as Visual-RFT (Liu et al. 2025) and VLM-R1 (Liu et al. 2025) successfully extended this approach to fundamental visual perception tasks; by designing verifiable rewards (e.g., IoU) and building general training frameworks, they validated its feasibility and revealed potential challenges like reward hacking. Building on this foundation, subsequent work has tackled more complex scenarios. UniVG-R1 (Bai et al. 2025a), for example, focuses on multi-image universal grounding problems by employing a 'cold-start' fine-tuning and a difficulty-aware strategy. In contrast, ViGoRL (Sarch et al. 2025), proposed in this paper, introduces a more fundamental mechanism by explicitly anchoring each reasoning step to specific image coordinates and introducing an innovative multi-turn RL framework, aiming to instill a deeper cognitive and interactive ability that is intrinsically tied to visual evidence.

3 GGBench: A Multi-Game Grounding Benchmark for LVLMs

This section mainly introduces the composition of GGBench, and the overall content is shown in Figure. 3. Section 3.1 elaborates on the design principles and task settings of GGBench. Section 3.2 presents the data collection process of GGBench. Section 3.3 explains the evaluation metrics of GGBench.

3.1 Design Principles and Task Formulation

Existing evaluation tools struggle to meet the complex demands of modern games: traditional visual benchmarks rely on real-world images that differ from game rendering styles, while legacy gaming benchmarks (e.g., Atari) fail to reflect high-fidelity graphics, complex UIs, and interaction logic of contemporary titles. To address this, GGBench was designed with core principles to ensure comprehensive and precise evaluation. It includes **104** representative modern mainstream games across genres like FPS and RPG to test generalization, and focuses on foundational skills—object grounding and UI element grounding—instead of ambiguous metrics like win/loss rates, resulting in a **1335-image** test set that offers actionable guidance for AI gaming.

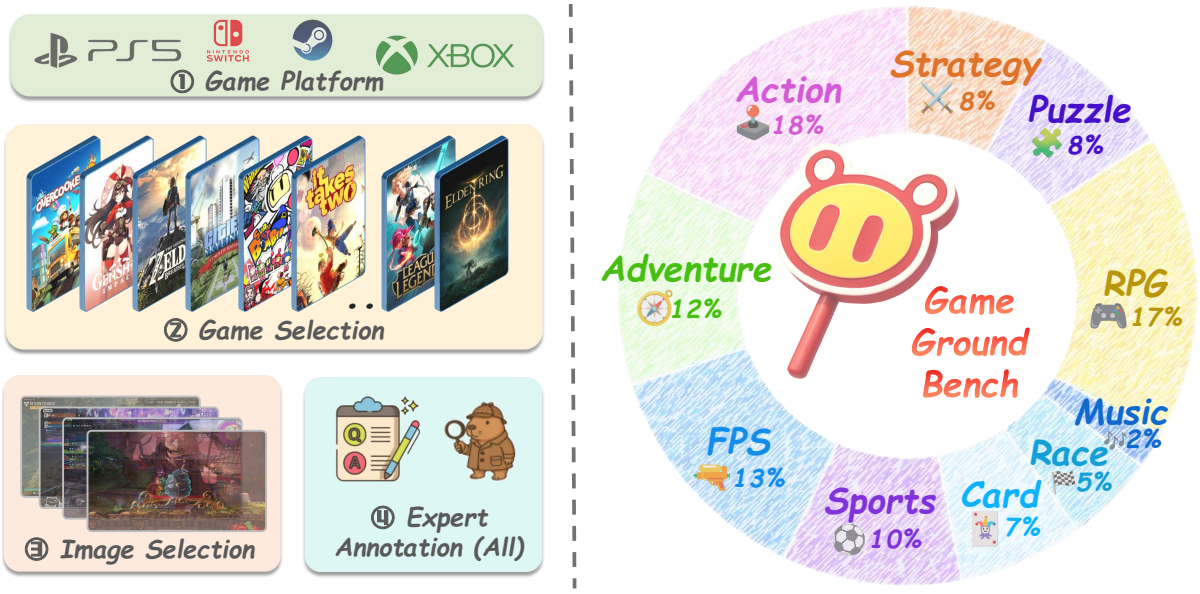

Figure 3: **The production process and composition diagram of GGBench.** GGBench selects images from popular games nowadays and conducts full manual annotation. It contains a total of **1335** images, covering **10** major game categories.

GGBench formulates visual grounding as a unified task: model f maps natural language queries T to bounding boxes B in game images I, aligning with in-game player intentions. The challenge of this task lies in the diverse range of text queries, which include both indirect object grounding for scene perception (e.g., '*Find the location of the enemy whose health is one-third less than the current target*') and indirect UI element grounding for interactive intent (e.g., '*Click the button that allows viewing the area where the user currently ranked second is located*'). Additionally, it incorporates indirect hybrid cases (e.g., '*Display the avatar of the teammate who ranks third in the number of rare items held*') to realistically evaluate the ability of context-dependent and robust grounding in complex visual environments.

3.2 Data Collection and Annotation

Game Platform and Selection. We first carefully selected the most popular games among players from mainstream platforms such as PlayStation 5, Xbox, and Nintendo Switch. On this basis, we further covered ten major game genres, with the proportion of each category as shown in Figure. 3. Among them are many popular contemporary games with complex visual effects, such as *Cyberpunk 2077*, *The Legend of Zelda*, and *It Takes Two*. This selection has built a rich test platform, which covers a wide range of gameplay mechanics and scene challenges, ranging from complex user interface management to navigation in high-fidelity open-world environments in modern role-playing games. In addition, to ensure that the model does not lean towards a single aesthetic style, the benchmark test has diversity in themes, including a variety of art styles such as cartoon, fantasy, sci-fi, and pixel art. This breadth of themes can effectively test the robustness of the model under different rendering techniques and visual languages.

Image Selection. To distill a high-quality and challenging dataset from this vast collection, we employed a rigorous, two-stage curation pipeline. **First**, we utilized a powerful Large Vision-Language Model (LVLM), Gemini-2.5-pro (Comanici et al. 2025), for an initial automated screening. The model was prompted to filter the archive for images that were both high-resolution in general and information-rich, prioritizing visually complex scenes or dense UI layouts. This automated process yielded a candidate set of high-fidelity images, the majority of which are at a resolution of 1920x1080 pixels. **Second**, our team meticulously reviewed each candidate image to ensure it was free of compression artifacts, representative of a meaningful gameplay moment, and presented a clear grounding challenge. This hybrid human-AI curation protocol allowed us to efficiently construct the visually diverse and high-quality dataset that forms the foundation of GGBench.

Annotation Protocol. The quality and challenge of GGBench stem from a rigorous manual annotation protocol. Trained annotators, familiar with each game genre, crafted diverse natural language queries (T) that move beyond simple descriptions to reference targets by their appearance, spatial relationships, or in-game functions, while also drawing tight bounding boxes (B). To ensure reliability, all annotations underwent a stringent two-stage review process involving both peer and expert checks. This meticulous protocol establishes GGBench as a high-quality and reliable benchmark.

A key feature of our protocol is its focus on queries requiring **indirect, multi-step** reasoning—unlike traditional localization tasks relying on immediate visual attributes, ours demand understanding of an object's function, state, or in-game context to deduce its location, mirroring human play-

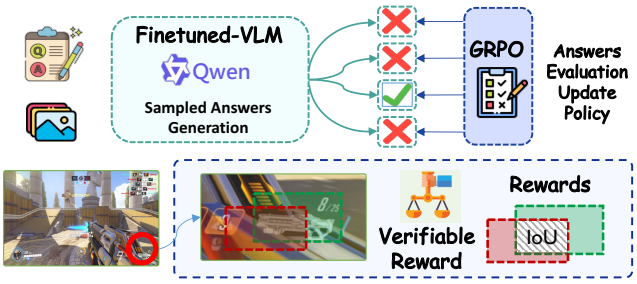

Figure 4: **Workflow of Game-R1.** This illustrates oGame-R1 trained with GRPO. Using IoU as a verifiable reward, this framework is suitable for fine-tuning on small-batch data, enabling rapid adaptation to new game content.

ers' cognition. Take the first example in the Figure. 2 to explain, it is the racing game ***F1 2025***, consider the question: '*Where is the Ferrari team's car that is closest to me?*'. In racing games, car positions depend on current rankings. The LVLM must first check the in-game rankings, find Ferrari drivers' positions, compare them with its own rank to identify the closest Ferrari, and lock onto it. This can't be done via visual traits alone; the model must grasp the game's ranking mechanics. This reasoning-centric protocol lets GG-Bench effectively test LVLMs' advanced localization abilities, making it a challenging, realistic benchmark.

3.3 Evaluation Metrics

To quantitatively assess model performance on this benchmark, we adopt the standard metric used in grounding and detection tasks: **Accuracy@IoU**. The core of this metric is the Intersection over Union (IoU), which measures the overlap between the predicted bounding box (B_{pred}) and the ground-truth bounding box (B_{gt}). A prediction is considered a correct localization if the IoU score exceeds a predefined threshold τ. Following established conventions, we set the primary threshold at $\tau = 0.5$. Therefore, the main reported metric throughout this paper is **Accuracy@0.5**, often denoted as Acc@50, which represents the percentage of test queries for which the model provides a correct localization.

4 Game-R1: Few-Shot Grounding across Universal Games

Our choice of training methodology is fundamentally driven by the unique and evolving challenges in the gaming domain. Unlike static benchmarks, the video game world is constantly changing, with new titles, genres, and interfaces emerging endlessly. Moreover, there is currently no unified large-scale platform for game data collection, making it too costly and impractical to create large-scale fully annotated datasets for each new game. This necessitates a paradigm adept at few-shot learning, enabling an agent to generalize well in new games or environments after training on a small amount of data. Thus, we built **Game-R1** based on the principles of Reinforcement Learning with Verifiable Rewards (RLVR), specifically using the Group Relative Policy Optimization (GRPO) algorithm. This R1-style approach is well-suited for the problem, as it distills robust policies from limited reward-based interactions rather than relying on massive supervised datasets.

The core of Game-R1 is the **Grounded Reinforcement Policy Optimization (GRPO)** (Shao et al. 2024; Guo, Zhang et al. 2025) algorithm, a data-efficient, online policy improvement method. As shown in Figure 4, the central idea of GRPO is to have the model conduct "group study": for a given localization problem, it generates (G) candidate predicted bounding boxes in parallel. Subsequently, a reward (r_i) is obtained by calculating the Intersection over Union (IoU) between each predicted box and the ground-truth box. Crucially, GRPO does not use these raw rewards directly. Instead, it calculates a normalized "advantage" score (A_i) using the following formula to evaluate each solution's relative performance within its group:

$$A_i = \frac{r_i - \mu r_1, \ldots, r_G}{\sigma r_1, \ldots, r_G} \quad (1)$$

Solutions above the group average get a positive advantage, those below a negative one. The policy (π_θ) is updated using these scores, favoring top-reward solutions and penalizing low-scoring ones. This online, critic-free internal comparison boosts data efficiency, outperforming traditional SFT significantly in few-shot scenarios like gaming.

5 Experiments

In this section, we mainly introduce the experimental part.

5.1 Implementation Details.

To establish a comprehensive leaderboard for GameGBench, we evaluate a wide array of state-of-the-art large vision-language models (LVLMs), including proprietary systems such as GPT-4o (OpenAI 2024) and Claude-3.5-Sonnet (Anthropic 2024), as well as open-source models from the LLaVA-NeXT (Anthropic 2024) and InternVL2.5 (Anthropic 2024) families. All open-source baseline models are fine-tuned on the full GGB training set using a standard supervised fine-tuning (SFT) recipe. Given that optical character recognition (OCR) capability is crucial for a benchmark focused on game screenshots, we specifically aim to examine whether Game-R1 can maintain its performance on OCR tasks (including OCRBench (Liu et al. 2024b) and MMVet (Yu et al. 2024)) and visual question answering (VQA) tasks (including MMBench (Liu et al. 2024a) and MMMU (Yue, Ni et al. 2024)) after undergoing SFT.

In contrast, our proposed Game-R1, built on the Qwen2.5-VL (7B) architecture, is **designed for extreme data efficiency**; its entire reinforcement learning phase utilizes an extremely small training set of only 500 images and is trained for just 500 steps. The primary evaluation of all models is conducted on the GGB test set using the ***Acc@50*** metric, as defined previously. Furthermore, to assess cross-domain generalization, the Game-R1 agent is also evaluated on the established RefCOCO/+/g (Kazemzadeh et al. 2014) real-world grounding dataset. All experiments were conducted on 8 NVIDIA A100 GPUs.

GGBench Leaderboard	Performance Across Game Genres (Acc@50)										Avg.
	Action	Adven	FPS	Sports	Card	Race	Music	RPG	Puzzle	Strategy	
Closed-source LVLMs											
GPT-4o (OpenAI 2024)	42.3	44.2	41.8	39.2	47.6	38.5	31.7	49.2	51.4	43.8	42.5
GPT-4.5 (OpenAI 2025)	31.5	34.2	29.3	25.9	36.1	28.4	23.7	36.3	39.5	32.8	31.8
Claude-3.5-Sonnet (Anthropic 2024)	40.1	43.5	37.2	33.4	45.2	36.6	30.1	46.8	49.2	41.5	40.4
Gemini-2.5-Pro (Comanici et al. 2025)	44.6	47.1	41.9	37.5	**49.3**	**40.2**	33.2	50.8	53.1	45.3	44.3
Grok-3 (xAI 2025)	29.4	31.6	27.1	24.1	33.5	26.3	22.2	33.9	36.8	30.3	29.5
HunYuan-Standard-V (Tencent 2024)	38.2	41.3	35.6	32.1	43.1	34.9	28.9	44.7	47.5	39.7	38.6
Open-source LVLMs											
LLaVA-NeXT-7B	27.3	29.3	25.3	23.1	31.7	24.8	20.9	32.3	35.4	28.6	27.9
LLaVA-OneVision-0.5B (Li et al. 2024a)	17.2	19.0	15.9	13.9	20.8	15.5	12.8	21.2	23.4	18.3	17.8
LLaVA-OneVision-7B (Li et al. 2024a)	28.8	30.8	26.7	24.2	33.2	26.2	22.3	33.7	36.8	30.1	29.3
LLaVA-OneVision-72B (Li et al. 2024a)	35.1	37.6	33.5	30.6	40.0	32.1	27.5	40.4	43.1	36.6	35.7
Ovis2-16B (Lu et al. 2024)	30.3	32.3	27.7	25.7	34.7	27.7	23.7	35.1	38.3	31.6	30.2
InternVL2.5-8B (OpenGVLab 2024)	32.3	34.3	29.7	27.6	36.6	29.6	25.6	37.0	40.2	33.5	32.6
Qwen2-VL-7B	31.1	33.1	28.5	26.3	35.5	28.6	24.5	36.2	39.3	32.4	31.6
Qwen2.5-VL-3B (Bai et al. 2025b)	27.2	28.2	25.8	24.7	30.4	24.1	23.5	31.0	33.7	27.6	27.6
Qwen2.5-VL-7B (Bai et al. 2025b)	33.4	35.5	30.8	28.6	37.6	30.7	26.6	38.3	41.2	34.8	33.8
Qwen2.5-VL-72B (Bai et al. 2025b)	38.5	40.6	35.9	33.4	42.3	35.2	29.6	44.2	46.8	39.8	38.6
Game-R1-3B	37.4	41.3	37.7	36.2	40.2	27.1	27.5	46.0	43.0	41.8	37.8
Game-R1-7B	**45.8**	**52.0**	**43.4**	**42.7**	45.0	37.9	**36.5**	**51.0**	**56.5**	**48.6**	45.9

Table 1: **Main Results of GGBench.** Using blue (darker for higher scores) for sub-item scores and purple for final averages. Closed-source models outperform open-source ones overall, with weaker performance in Sports, Music and Strategy. Our GRPO and simple R1 fine-tuning significantly boost performance and balance metrics across categories.

5.2 Main Results

First, we examine the experimental results of various models on our GGBench. Then, we test the changes in VQA and OCR capabilities of Game-R1 after the fine-tuning.

GGBench Leaderboard. The GGB Leaderboard (Table 1) and radar figure (Fig. 1) reveals additional insights:

- Closed-source LVLMs generally outperform open-source ones, but the open-source Game-R1-7B stands out by surpassing all closed-source models, highlighting the potential of game-specific fine-tuning.
- Performance varies across genres: Puzzle and RPG yield higher scores, while Music and Racing lag behind.
- Larger open-source models outperform smaller counterparts but still trail Game-R1-7B, demonstrating that the Reinforcement Fine-tuning is effective.
- Game-R1-7B excels in multiple genres, with Gemini-2.5-Pro leading in Card and Racing, while closed-source models show more consistent cross-genre performance than most open-source alternatives.

RefCOCO/+/g. We compare several Large Vision-Language Model (LVLM) methods used for grounding, including KOSMOS-2 (Peng et al. 2024), Shikra-7B (Chen et al. 2023), VisionLLM-H (Wang et al. 2023), and others. As shown in Table 2 for the RefCOCO, RefCOCO+, and RefCOCOg benchmarks, LVLM methods display notable performance differences across the various test sets. Among them, Game-R1 7B stands out, achieving the highest score of 90.1 on the RefCOCO validation set, demonstrating strong visual–language grounding capability. Despite being trained exclusively on game data, Game-R1 also performs remarkably well on real-world benchmarks, confirming its strong generalization ability. This shows that grounding skills learned from game environments transfer effectively to real-world tasks, validating the effectiveness of our training scheme.

VQA and OCR Bench. VQA enables models to interpret in-game visuals, while OCR converts on-screen text into usable form; together, they are essential for understanding game mechanics, narratives, and player needs. We therefore aimed to check whether GGB finetuning harms performance in these areas. As shown in Table 3, Game-R1 maintains excellent VQA and OCR results after finetuning. On MM-Bench (MM-B) and MMMU, the 3B model scores 77.1 and 52.0, while the 7B model reaches 82.8 and 58.9—strong results for their parameter sizes. On OCRBench (OCR-B), Game-R1 3B matches Qwen2.5-VL 3B at 828, and the 7B model scores 884, close to Qwen2.5-VL 7B. The 7B model also leads MMVet with 69.9, confirming that the model re-

Grounding	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
LVLM-based method								
○ KOSMOS-2	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.6
○ Shikra-7B	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
○ VisionLLM-H	86.7	86.7	-	-	-	-	-	-
• Qwen2.5-VL 3B	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7
• **Game-R1 3B**	88.6	91.2	84.6	82.3	88.3	75.2	85.9	85.2
• Qwen2.5-VL 7B	90.0	**92.5**	**85.4**	84.2	89.1	76.9	87.2	**87.2**
• **Game-R1 7B**	**90.1**	92.3	84.7	**84.3**	**89.6**	**77.3**	**88.1**	86.6

Table 2: **Grounding Results of RefCOCO series.**

Understanding	VQA		OCR	
	MM-B	MMMU	OCR-B	MMVet
Open-source LVLMs				
○ LLaVA-OneVision-0.5B	56.8	32.7	583	31.5
○ LLaVA-NeXT-7B	63.0	35.8	532	40.2
○ LLaVA-OneVision-7B	76.8	48.8	622	51.9
• Qwen2.5-VL 3B	76.8	53.1	828	60.0
• **Game-R1 3B**	77.1	52.0	828	58.7
• Qwen2.5-VL 7B	82.2	58.6	**888**	69.7
• **Game-R1 7B**	**82.8**	**58.9**	884	**69.9**

Table 3: **Understanding Results of VQA and OCR.**

SFT *vs* RFT	GGBench		
	FPS	Card	Race
Game-R1 7B			
• SFT 300 steps	34.7	39.7	31.7
• RFT 300 steps	41.2	43.4	34.7
• SFT 500 steps	39.4	42.4	35.7
• RFT 500 steps	**43.4**	**45.0**	**37.9**

(a) **Supervised / Reinforcement Fine-Tuning**

Reward	GGBench		
	FPS	Card	Race
Game-R1 7B			
• IoU ≥ 0.3	32.4	41.7	34.0
• IoU ≥ 0.5	35.4	43.8	36.0
• IoU ≥ 0.7	34.7	42.8	35.5
• IoU Soft	**43.4**	**45.0**	**37.9**

(b) **The Verifiable ReWard in GRPO**

Cross-Genre Training Data	GGBench		
	FPS	Card	Race
Game-R1 7B	**43.4**	**45.0**	**37.9**
• FPS only	**45.7**	40.7	32.9
• Card only	36.8	**46.4**	32.5
• Race only	37.8	40.8	**38.7**
• Music only	36.4	41.1	33.2

(c) **Cross-Genre Training Data**

Table 4: **Ablation Study of Game-R1 on GGBench.** Here, • represents the method used in our Game-R1 series.

tains strong VQA and OCR capabilities post-finetuning.

5.3 Ablation Studies

Efficacy of the GRPO Stage. We first isolate the contribution of the GRPO-based reinforcement learning stage. We compare the full Game-R1 model with a version trained only via supervised fine-tuning (SFT) on the same data. As shown in Table 4a, after 500 training steps, the full model reaches an average score of 42.1 across the three genres, outperforming the SFT-only model's 39.2. This indicates that on a diverse but limited dataset, simple supervised imitation is insufficient, and the exploration and optimization provided by GRPO are essential for learning a robust policy.

Impact of the Reward Function. We examine the design of our reward signal by comparing our standard model, which uses a continuous IoU score as a soft reward, against variants with discrete threshold-based rewards. As shown in Table 4b, the soft IoU reward delivers the best performance, reaching an average score of 42.1. In contrast, the best discrete variant—using a binary reward for IoU ≥ 0.5—scores only 38.4. This drop indicates that fine-grained, continuous feedback provides a more effective learning signal, allowing GRPO to perform more precise policy updates.

Impact of Cross-Genre Training Data. To verify the importance of data diversity, we evaluate "specialist" agents trained We also evaluate models trained exclusively on a single game genre. As shown in Table 4c, these specialists perform well on their own genre—for example, the FPS-only agent scores 45.7 on FPS, exceeding the standard model's 43.4—but their performance collapses on out-of-genre tasks. The FPS-only agent drops to 40.7 on Card and 32.9 on Race, yielding a lower overall average of 39.8 compared to the standard model's 42.1. This underscores that cross-genre training data is crucial for learning a truly generalizable grounding policy.

6 Conclusion and Limitation

We introduce GGBench, a large-scale cross-genre benchmark designed to probe the limits of LVLMs' semantic grounding in modern game universes. To address data-efficient learning in this domain, we propose Game-R1, a training method based on the GRPO algorithm. With only 500 fine-tuning samples, Game-R1 surpasses all open-source models on GGBench and even outperforms leading closed-source systems, while showing strong generalization to real-world tasks. Despite these gains, our work has limitations: the benchmark uses static screenshots rather than dynamic gameplay, and the grounding task is limited to bounding box prediction. Future work will extend the benchmark to video and incorporate richer interaction models to better support the development of gaming agents.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62422606, 62201484).

References

Anthropic. 2024. Claude 3.5 Sonnet.

Bai, S.; Li, M.; Liu, Y.; Tang, J.; Zhang, H.; Sun, L.; Chu, X.; and Tang, Y. 2025a. UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning. *arXiv:2505.14231*.

Bai, S.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv:2502.13923*.

Baker, B.; Akkaya, I.; Zhokhov, P.; Huizinga, J.; Tang, J.; Ecoffet, A.; Houghton, B.; Sampedro, R.; and Clune, J. 2022. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos. In *NeurIPS*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Cheng, K.; Sun, Q.; Chu, Y.; Xu, F.; Li, Y.; Zhang, J.; and Wu, Z. 2024. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. In *ACL*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Garrette, D.; Luan, D.; Petrov, S.; and Kavukcuoglu, K. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Guo, D.; Zhang, Z.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. VTimeLLM: Empower LLM to Grasp Video Moments. In *CVPR*.

Hudi, F.; Winata, G. I.; Zhang, R.; and Aji, A. F. 2025. TextGames: Learning to Self-Play Text-Based Puzzle Games via Language Model Reasoning. *arXiv:2502.18431*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Tu, V.; et al. 2024b. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Liu, Y.; Duan, H.; Xu, Y.; Wei, H.; Song, Z.; Zhang, J.; Li, Y.; Li, Z.; Xie, W.; Fan, J.; and Dong, H. 2024a. MMBench: Is Your Multi-modal Model an All-around Player? In *International Conference on Learning Representations*.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.-C.; Liu, C.-L.; Jin, L.; and Bai, X. 2024b. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-RFT: Visual Reinforcement Fine-Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv:2405.20797*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*.

OpenAI. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680*.

OpenAI. 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-02-15.

OpenAI. 2025. OpenAI GPT-4.5 System Card. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf. Accessed: 2025-02-15.

OpenGVLab. 2024. InternVL 2.5: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. https://internvl.github.io/blog/2024-12-05-InternVL-2.5/.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2024. Kosmos-2: Grounding Multimodal Large Language Models to the World. In *International Conference on Learning Representations*.

Qin, Y.; Ye, Y.; Fang, J.; Wang, H.; Liang, S.; Tian, S.; Zhang, J.; Li, J.; Li, Y.; Huang, S.; Zhong, W.; Li, K.; Yang, J.; Miao, Y.; Lin, W.; Liu, L.; Jiang, X.; Ma, Q.; Li, J.; Xiao, X.; Cai, K.; Li, C.; Zheng, Y.; Jin, C.; Li, C.; Zhou, X.; Wang, M.; Chen, H.; Li, Z.; Yang, H.; Liu, H.; Lin, F.; Peng, T.; Liu, X.; and Shi, G. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv:2501.12326*.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *CVPR*.

Sarch, G.; Saha, S.; Khandelwal, N.; Jain, A.; Tarr, M. J.; Kumar, A.; and Fragkiadaki, K. 2025. Grounded Reinforcement Learning for Visual Reasoning. In *NeurIPS*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Wang, W.; Guo, D.; Zhu, Q.; and Dong, X. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Tencent. 2024. Tencent Hunyuan.

Tsai, C. F.; Zhou, X.; Liu, S. S.; Li, J.; Yu, M.; and Mei, H. 2023. Can large language models play text games well? current state-of-the-art and open questions. *arXiv preprint arXiv:2304.02868*.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.;

Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gulcehre, C.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Wang, W.; Xu, J.; Dai, J.; Liu, Z.; Hu, X.; Li, L.; Li, R.; Shi, P.; Wang, Y.; Li, Y.; and Luo, P. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, volume 36.

xAI. 2025. Grok 3 Beta — The Age of Reasoning Agents.

Yu, P.; Li, Z.; Zhang, Z.; Zhang, Y.; Wang, R.-Z.; Liu, Z.; Cui, A. Y.; Xu, Z.; Zhu, Y.; Shi, X.; Li, M.; and Smola, A. 2025. RPGBench: Evaluating Large Language Models as Role-Playing Game Engines. *arXiv preprint arXiv:2502.00595*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *ICML*.

Yue, X.; Ni, Y.; et al. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*.

Zhang, Y.; Zhang, P.; Wu, L.; Song, Y.; Liu, H.; Yu, J.; and Liu, Z. 2024. NExT-Chat: An LMM for Chat, Detection and Segmentation. In *Proceedings of the 41st International Conference on Machine Learning*.

Zhao, W.; Lin, Z.; Zhou, D.; Huang, Z.; and Kang, B. 2023. BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs. *arXiv preprint arXiv:2307.08581*.