Vision vs. Text: Quantifying Multimodal Performance in Embodied Reasoning and Long-Horizon Planning

Kaiden Jang

January 31, 2026

Abstract

As large language models (LLMs) transition from passive virtual assistants to active embodied agents, understanding their performance in 3D settings is becoming increasingly critical. This research project investigates the performance disparities in state-of-the-art multimodal LLMs between vision and text-based inputs in complex, 3D embodied environments. It utilizes the open-ended nature of Minecraft with the MINDcraft framework to evaluate three modern LLMs: Claude Opus 4.5, Gemini Robotics-ER 1.5, and GPT-5.2. These models are evaluated through material progression, unique item collection, and number of deaths. The experiment finds that 3D-environment proficiency with both vision and text inputs varies greatly between different model architectures. Specialized Vision-Language-Action models like Gemini Robotics excel with specific embodied and spatial reasoning using vision inputs. Other abstract reasoning models dominate high-level reasoning but lack low-level multimodal capabilities. The study concludes that ideal 3D-environment LLM performance requires an optimization of both specialized low-level processes while maintaining high abstract reasoning capabilities.

1 Introduction

While the future seems to be becoming increasingly dependent on Large Language Models (LLMs), LLMs may not be as capable as people perceive them to be. The past five years have brought great advancements to LLMs with newer, more advanced models constantly breaking previous benchmarks. While LLMs originally functioned as passive virtual assistants or chatbots, their capabilities are evolving beyond simple text processing. Advancements are allowing these models to interact directly with environmental stimuli, bridging the gap between a text-based assistant to an active embodied agent

1.1 Research Context

Recently, a new branch of LLM research has been exploring their ability to operate in 3D environments. Specifically, research has focused on LLMs' capacity for embodied reasoning and long-horizon planning. Embodied reasoning is the ability for LLMs to use perception, action, and environmental feedback to ground their thoughts in the real world. Long-horizon planning refers to the ability for LLMs to create and execute plans that require many steps. These abilities are crucial for effective operation in 3D environments. However, research has also found that LLMs typically fall short in both these areas, typically due to something called model hallucination. Models are known to hallucinate or generate information that does not exist. In the context of embodiment and long-horizon planning, hallucinations can cause models to get stuck in logic loops, making them a major issue in LLM use in 3D environments.

Effective research and testing model embodiment and long-horizon planning requires complex, open-ended 3D environments that can nurture creative reasoning. For this, Minecraft fits the requirements due to its open-ended nature as a sandbox game. This environment allows for the evaluation of embodied LLMs that navigate and interact with the world through in-game characters. There are an infinite number of approaches towards playing the game. Despite the provided progression track of material advancements, which involves gathering resources more advanced resources to craft better tools, the game leaves room for creativity regarding how to achieve these advancements. This open-ended nature provides us with the perfect testbed for long-horizon planning in models.

When interacting with 3D environments like Minecraft, most current model agents do not interact with their environment like a human would. Most agents are text-based. This means they receive text-based descriptions of their surroundings instead of actually “seeing” them. The lack of vision-based agents in this field is partially due to LLMs’ initially only accepting text inputs. However, recent advancements in LLMs provide us with highly capable multimodal language models. These models allow both visual and textual information to be inputted. Current state-of-the-art models are primarily multimodal large language models (MLLMs). While text-based agents often miss finer details in embodied scenarios, multimodal agents have the potential to overcome these limitations. Direct visual inputs in multimodal models also have the potential to minimize hallucinations since models are provided with direct data of their environment.

Advancements in LLMs have also led to specialized models for embodied reasoning, such as Gemini Robotics-ER 1.5. This model is part of an emerging field of models explicitly designed for interaction with physical environments. These specialized models have a unique Vision-Language-Action (VLA) architecture and are designed to mitigate the spatial and embodied reasoning limitations of traditional LLMs. VLA models can greatly improve performance in these environments even with less general reasoning capabilities.

1.2 Research Objective

Theoretically, multimodal models could offer better agent performance in 3D environments like Minecraft through superior grounding by means of visual inputs, decreasing hallucination, and increasing long-horizon and embodied reasoning capabilities. However, the extent to which visual perception improves performance over traditional text-based inputs remains largely unexamined. This research aims to quantify the difference between vision-based agents and text-only agents by directly comparing the embodied reasoning and long-horizon planning performance of the two in state-of-the-art VLA and general multimodal models utilizing the open-endedness of Minecraft.

1.3 Hypothesis

This study is guided by two primary hypotheses concerning LLM performance in embodied 3D environments. First, it is hypothesized that vision-based agents will perform more proficiently in such embodied environments compared to their text-only counterparts. The direct visual perception will demonstrate detailed data to the LLMs that help provide the environmental grounding necessary for high performance in embodied environments. Second, it is hypothesized that specialized VLA architecture models like Gemini Robotics-ER 1.5 will show similar or superior performance than that of larger general-purpose models such as GPT-5.2 and Claude Opus 4.5 with visual inputs. While these general-purpose reasoning models possess higher levels of reasoning power, Gemini's specialized VLA architecture is expected to offer superior embodied reasoning performance with visual inputs.

2 Related Works

Minecraft for Embodied AI Research. Since it is a vast, open-ended embodied world with complex dynamics, Minecraft has become a popular tool for evaluating embodied reasoning and long-horizon planning capabilities of LLM-based agents. Early work, such as that of the Voyager project led by Nvidia, in collaboration with researchers from the California Institute of Technology and other leading institutions, was the first to utilize Minecraft's embodied environment to test LLMs. Building on this, researchers at Heriot-Watt University introduced VoyagerVision which expanded the limitations of Voyager by implementing vision inputs to greatly increase agents' capabilities to create specific structures. Research using embodied agents within Minecraft was further expanded with the development MINDcraft an embodied, multi-

agent collaborative framework by Isadora White and her team at the University of California, San Diego. This new framework provides a versatile platform for the testing of the collaborative and embodied capabilities of LLMs.

Multimodal Architecture and Efficiency. The architecture and token efficiency of MLLMs is another important area of research when evaluating multimodal agents. In this context, tokens are the basic units of information that a model processes. Token efficiency refers to the ability to achieve high performance while minimizing the token count. Works such as the study on "Multi-Stage Vision Token Dropping" (MustDrop) by Ting Liu et al. at the National University of Defense Technology highlighted the inefficiency of most visual tokens by finding that most are redundant information. However, other works such as Yanhong Li and her team's study on Evaluating Efficiency and Understanding of LLMs with Visual Text Inputs at Allen Institute for AI find structured visual tokens are more efficient than text token at understanding textual data under certain conditions. A survey by Shukang Yin and a team of researchers details the three-part encoder-connecter-LLM architecture that maps how models align different modalities into a unified framework. Additionally, specialized vision-language-action (VLA) architecture models like Gemini Robotics-ER 1.5 have established an unprecedented standard for embodied and spatial reasoning. Gemini Robotics and Google DeepMind's paper on the model explains that this specialized VLA structure reduces planning failures that can outperform general-purpose models in physical reasoning.

3 Methodology

3.1 Environment and System Parameters

The experimental 3D environment was configured in Minecraft version 1.21.1. Minecraft was selected as the environment due to its open-ended nature and complex dynamics that can effectively evaluate models' long-horizon planning and embodiment capabilities. Version 1.21.1 was chosen primarily because it is the most stable version supported by the MINDcraft framework, which was utilized to facilitate model communication within the grounded environment of Minecraft. The latest release of MINDcraft(Mindcraft v0.1.3 | Villagers and Buckets) was used. The MINDcraft framework managed environmental information and action execution, translating the LLMs' outputs into actionable commands.

To ensure consistent performance across all trials, MINDcraft was configured with specific parameters. The *allow_insecure_coding* feature was enabled, allowing all models to both call upon pre-existing commands and generate custom commands. MINDcraft handles this by taking in a natural language input following the *!newAction* command. MINDcraft then transforms the natural language input into a code generation prompt and sends it to another model to generate the actual code. However, a unique code generation model was not specified in MINDcraft, and all code generation requests were handled by the LLM used in each run. Additionally,

the *!endGoal* command was listed as a blocked action to prevent models from autonomously ending the current task prematurely.

3.2 Model Selection and Agent Profiles

This study evaluated three distinct models: GPT-5.2, Claude Opus 4.5, and Gemini Robotics-ER 1.5. Each model was configured to utilize two different perceptual modalities: text inputs and visual inputs. This totaled six distinct profiles, all configured within MINDcraft to test differences between the embodied reasoning and long-horizon planning capabilities of three cutting-edge LLMs. Within MINDcraft, for the text-based agent trials, the model being tested was setup in its MINDcraft agent profile. Visual agents, however, require a model for processing images and the main agent loop allowing for non-multimodal models to be used. To maintain experimental integrity, only multimodal models were selected for the experiment, allowing for a single model to handle both image processing and the main agent loop. GPT-5.2 and Claude Opus 4.5 were selected as they are the current flagship models of leading AI companies OpenAI and Anthropic, respectively. Both models are praised for their general reasoning, with Claude Opus 4.5 particularly being known for its code generation abilities. Gemini Robotics-ER 1.5 is one of the latest releases from Google Gemini family. This model was built around a specialized Vision-Language-Action (VLA) architecture that allows it to excel in embodied reasoning tasks, especially when utilizing vision inputs. This unique specialization is why it was chosen for the experiment. By incorporating all six profiles, this experiment can isolate the impacts of visual perception on model performance across both general-purpose models and specialized VLA architectures.

3.3 Data Collection and Metrics

To automate the data collection process, a custom script was developed by the author for extracting Minecraft JSON statistics[1]. This script was used to extract data from each agent's Minecraft stats file, which were used to compare performance across all agent profiles. The data collection script organizes information into three primary metrics to evaluate embodied reasoning and long-horizon planning: *Unique Item Collection* (measuring the distinct items picked up, crafted, or used), *Death Count*, and *Material Progression* (the hierarchical resource tier upgrade path). These metrics effectively reflect the performance capabilities of models, allowing for evaluation of long-horizon planning and embodied reasoning within their environment. For example, a higher material progression tier can mean more advanced long-horizon planning capabilities as it suggests an agent was able to effectively plan future steps that allowed for quicker progression through the tiers. Higher performance in the *Unique Item Collection* metric corresponds with higher long-horizon planning capabilities as effective planning would avoid too many repeat actions that collect the same item. A lower number of

[1] https://github.com/GGaego/mllm-minecraft-agent-stats

deaths would suggest an application of higher spatial or embodied reasoning capabilities needed to avoid environmental hazards.

Tier Level	Resource Requirement
Tier 0	Wood Logs
Tier 1	Wood Tools
Tier 2	Stone
Tier 3	Stone Tools
Tier 4	Raw Iron
Tier 5	Iron Ingots
Tier 6	Iron Tools
Tier 7	Diamond
Tier 8	Diamond Tools
Tier 9	Obsidian

Table 1: Hierarchical Material Progression Tiers for quantifying Long-Horizon Planning

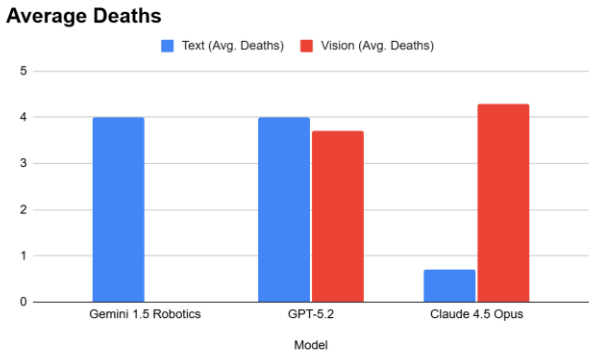
Material Progression Hierarchy. A hierarchical tier system ranging from 0-9 was created to measure each agent's material progression through the game. The tier system categorized materials by their difficulty to obtain, and the specific resource requirements for each tier are detailed in Table 1. Completion of each tier was measured by checking if the tool/resource was crafted or mined. This material progression hierarchy established a uniform way to numerically compare each agent's progression through the game. Since one of the goals of this experiment was to evaluate long-horizon planning, establishing this material progression hierarchy was essential.

3.4 Experimental Procedures

To maintain a consistent test environment, the environment was set within Minecraft version 1.21.1 with the seed -4029535714769340309. This seed was chosen due to an absence of structures in the starting "plains biome" area. The plains biome was chosen since it provided agents with a neutral environment containing resources that would allow them to naturally progress through the game. Structures were avoided to prevent agents from skipping tiers in the material progression hierarchy created. Each trial world was a copy of a base world which was set to -4029535714769340309 and had vanilla commands except for "allow commands," which was enabled.

Each trial began at time 0 in a fresh copy of the world. Agents were initialized in MINDcraft and loaded into the world, which was hosted locally through a LAN connection. After spawning into the world, they were prompted in chat with the goal command titled *!goal("play Minecraft, survive, and progress through the game")*. This *!goal* feature directs the agents to continuously prompt themselves to use actionable commands to interact with the world. This was repeated three times per unique agent profile.

4 Results



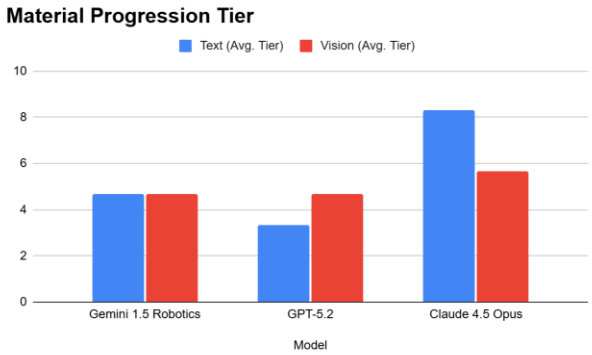


Figure 1: Average Number of Agent Deaths

Figure 2: Average Material Progression Tier

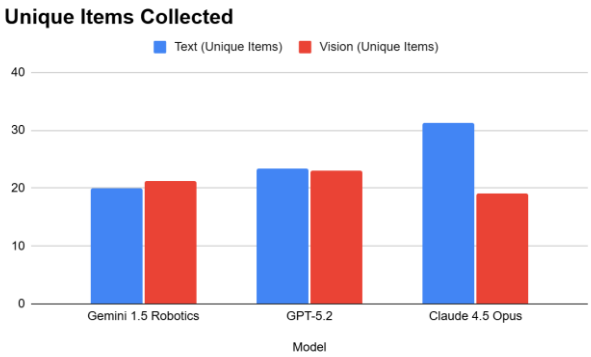


Figure 3: Average number of Unique Items Collected

Throughout the experiment, no clear performance gap between text-based and vision-based agents was observable. Some models performed better with one modality while some performed better with the other. Figures 1, 2, and 3 reveal that Claude Opus 4.5 performed significantly better with textual inputs across all three metrics. Claude's visual agents performed the worst compared to other models while its textual agents outperformed all other agents in all metrics. Conversely, while the margin was not as significant as Claude's, Gemini 1.5 Robotics models performed better with visual inputs in most tasks compared to textual inputs. GPT-5.2's data shows a smaller performance gap between both perceptual modalities compared to other models. GPT only showed a noticeably higher average material progression tier with visual agents.

Material Progression. Comparing the *Material Progression* results in Figure 2, model performance across both modalities was dependent on the model. Gemini's vision and text-based agents both scored exactly 4.67 in this metric. GPT-5.2 agents achieved a slightly higher average across visual agents (4.67 versus 3.33). Claude Opus 4.5 achieved the highest tier average among the text agents at 8.33, reaching tier 9 twice. Claude's visual agents, however, scored significantly lower than their textual counterparts, scoring almost 3 tiers lower.

Unique Item Collection. Looking at Figure 3, while GPT-5.2 and Gemini both performed similarly across both modalities, Claude saw significant differences between both modalities. Figure 3 shows that Claude Opus 4.5 text models collected significantly more unique items than any of GPT-5.2 and Gemini Robotics-ER 1.5 agents. Interestingly, Claude Opus 4.5 vision agents performed the worst in this metric, collecting the least number of unique items. GPT-5.2's text agents only performed slightly better than its visual agents. Gemini Robotics-ER 1.5 agents showed the opposite, with its visual agents collecting an average of 1 more unique item than its text agents.

Death Count. Figure 1 shows us that visual input agents generally incurred fewer deaths than text agents. The exception for this is Claude which conceded around 4 deaths with visual inputs versus the near zero deaths seen with its text agents. Conversely, Gemini had an average of 0 deaths in the vision trials but an average of 4 deaths in text trials. In this metric, GPT-5.2 had a less significant difference in deaths compared to other models, averaging around 4 deaths for each.

5 Discussion

While no consistent performance gap was observed between modalities across all three models, Claude Opus 4.5 showed a clear efficiency gap between its textual and visual agents. Gemini Robotics-ER 1.5 showed differences in specific metrics, while GPT-5.2 displayed almost no significant differences between textual and visual modalities. These distinct performance discrepancies between models can likely be attributed to the differences between model architecture and training data used to create these models. Specifically, the data suggests that different models are more suitable for long-horizon planning or embodied reasoning. When it comes to tasks requiring high-level abstract reasoning, such as material progression and long-horizon planning, Claude Opus 4.5 excels since it is an agentic model (models designed to autonomously complete complex goals) specifically tuned for high-level reasoning. Also, differences between each model's performance with vision and text inputs suggest model architecture and training objectives are important factors in their sensitivity to input modality. This is supported by the inconsistency between models and input modality.

Collectively, these results suggest a trade-off in current multimodal AI models between higher-level abstract reasoning and low-level image processing needed for spatial grounding and effective embodied reasoning. While advanced agentic models like Claude excel in high-level

abstract reasoning but suffer when integrating visual information, specialized VLA architecture models like Gemini Robotics prioritize visual interpretation, improving their environmental grounding and spatial reasoning. GPT-5.2 displayed more neutral results while still proving that efficient multimodal integration can be more efficient with embodied agents than textual data.

5.1 Claude Opus 4.5

Claude Opus 4.5 was notably more proficient with textual inputs across all metrics. This performance gap suggests a training focus on text rather than image processing capabilities. Furthermore, when using text inputs, Claude clearly presented itself as the most proficient model at the task. As the experiment is aimed at evaluating embodied reasoning and long-horizon planning capabilities, Claude likely has more advanced embodied reasoning and planning capabilities. Since Claude models have advanced coding abilities, which require high levels of reasoning and planning, it follows that Claude Opus 4.5 excelled in this testing environment.

The performance gap between modalities displayed by Claude Opus 4.5 suggests that its specialization in high-level logic can be a major limitation when tasked with low-level processes such as image processing or interpretation. While Claude performed reasonably well in all aspects of the experiment across modalities, its comparative struggle with visual inputs suggests inefficiencies in low-level, multimodal reasoning capabilities when incorporating visual information. One potential explanation is that rather than seamlessly integrating visual data like other models, Claude incurs significant computational overhead to the point that any benefits gained from the increased detail are counteracted. This inefficiency likely distracts the model from its main high-level reasoning or planning tasks. During the trials, instead of enhancing environmental awareness and grounding the model, it appears that the visual inputs add another layer of complexity which degraded its core reasoning abilities it typically excels in resulting in this clear performance gap.

5.2 Gemini Robotics-ER 1.5

While both modalities displayed similar material progression and number of unique items collected, Gemini's vision agents incurred no deaths during all three trials while its text agents had an average of 4 deaths per trial. *Material Progression* and U*nique Item Collection* metrics showed less variation between modalities. This is likely due to both metrics' focus on long-horizon planning capabilities as opposed to spatial reasoning, in which the model's specific VLA architecture excels. This also means the modality had less of an effect on long-horizon planning capabilities in Gemini likely because the high-level reasoning ability needed for long-horizon planning is not heavily affected by different modalities even with the VLA architecture.

In contrast to Claude, Gemini Robotics-ER 1.5's Vision-Language-Action (VLA) architecture allows the model to better incorporate visual information into 3D maneuvering. The results reveal that although metrics favoring high-level reasoning remained consistent, the number of

deaths was drastically reduced by visual inputs. Since death count is highly dependent on spatial or embodied reasoning (needed to effectively avoid environmental hazards), models with advanced multimodal reasoning perform significantly better in that metric. This suggests that, unlike Claude, Gemini's multimodal reasoning is advanced enough that it does not negatively impact the model's high-level reasoning abilities when synthesizing textual and visual information. It also suggests that its spatial awareness effectively grounds the model in the world, which boosts its spatial and embodied reasoning. These conclusions are supported by the fact that the Gemini Robotics model is specifically trained to interpret visual data and use advanced spatial reasoning to ground the model in the environment. Consequently, this specialized architecture provides the most logical explanation for the lower death count.

5.3 GPT-5.2

Unlike the other two models, GPT-5.2 showed smaller differences between modalities. The most significant difference found was a slightly higher average material progression tier with vision. However, the difference was only about 1 tier higher. The results for the other two metrics, D*eath Count* and U*nique Item Collection*, were almost identical. The higher average material progression for GPT's visual agents suggests a more advanced image processing system. This would reduce the type of information clutter that misaligns models from their long-term goals.

Unlike the other two models, GPT-5.2 showed near consistent performance across all three metrics. This suggests that the model's multimodal integration is seamlessly integrated to where it doesn't negatively impact performance. The data reveals a slight increase in material progression with visual inputs. From this, it can be inferred that GPT-5.2 possesses better multimodal integration that doesn't negatively impact performance with added complexity as seen in Claude. Seamless multimodal integration would mean being able to extract only necessary data from the image without distracting the model from high-level reasoning tasks. This indicates that GPT-5.2 has a balanced architecture effective for efficient multimodal integration where visual and textual data are equally viable for embodied and long-horizon, high-level reasoning.

5.4 Limitations

While the findings of the study provide significant insight into the multimodal capabilities of state-of-the-art LLMs in embodied and high-level planning tasks, several limitations need to be acknowledged. The main limitation of this research is the limited number of trials for each model and the input modality combination tested. While the results and performance patterns remained mostly consistent, more trials would be necessary for statistically robust results and to mitigate the variance between model outputs. Furthermore, the MINDcraft framework presented several specific challenges affecting agent stability during trials. These challenges included minor bugs causing agents to automatically restart, agents getting caught in loops, and hardcoded defense logic within agents. When put together with the small trial count, the bugs may have influenced

specific behaviors. Additionally, the hardcoded defense logic likely influenced death counts as models had no control over decisions to fight or flee from specific situations. Finally, it is important to note that the block-based world logic of Minecraft differs greatly from real world physics. Because of this, the results and conclusions observed in models may not translate directly to real world robotics or other similar studies.

6 Conclusion

This study investigated the performance differences between current state-of-the-art multimodal LLMs with text-based or vision-based agents within an embodied setting. The results show a clear trade-off between high-level abstract reasoning and low-level multimodal reasoning in modern LLMs. While models like Claude Opus 4.5 show degraded performance when incorporating visual inputs, models with specialized architectures for vision, like the VLA in Gemini Robotics-ER 1.5, excel at synthesizing textual and visual information. This VLA model demonstrated superior environmental grounding and spatial awareness through hazard avoidance. GPT-5.2 highlights neutrality between both modalities achieving mostly consistent results regardless of input modality. In conclusion, these findings suggest the future of autonomous embodied agents lies not within a specific modality such as text or vision, but on a balance between abstract reasoning power and low-level specialized multimodal capabilities. As AI moves towards real-world embodiments, this balance between high-level abstract planning and low-level visual and spatial understanding will be essential for creating agents intellectually and physically capable.

Works Cited

Li, Yanhong, et al. "Text or Pixels? Evaluating Efficiency and Understanding of LLMs with Visual Text Inputs." *Findings of the Association for Computational Linguistics: EMNLP 2025*, vol. Findings of the Association for Computational Linguistics: EMNLP 2025, Jan. 2025, pp. 10564–78, https://doi.org/10.18653/v1/2025.findings-emnlp.558.

Liu, Ting, et al. "Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model." *ArXiv.org*, 2024, arxiv.org/abs/2411.10803. Accessed 28 Jan. 2026.

Smyth, Ethan, and Alessandro Suglia. "VoyagerVision: Investigating the Role of Multi-Modal Information for Open-Ended Learning Systems." *ArXiv.org*, 2025, arxiv.org/abs/2507.00079. Accessed 28 Jan. 2026.

Team, Gemini Robotics, et al. "Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer." *ArXiv.org*, 2025, arxiv.org/abs/2510.03342.

Wang, Guanzhi, et al. "Voyager: An Open-Ended Embodied Agent with Large Language Models." *ArXiv.org*, 25 May 2023, https://doi.org/10.48550/arXiv.2305.16291.

White, Isadora, et al. "Collaborating Action by Action: A Multi-Agent LLM Framework for Embodied Reasoning." *ArXiv.org*, 2025, arxiv.org/abs/2504.17950. Accessed 28 Jan. 2026.

Yin, Shukang, et al. "A Survey on Multimodal Large Language Models." *ArXiv.org*, 23 June 2023, https://doi.org/10.48550/arXiv.2306.13549.